VI

Learning Effective Classifiers from Local Descriptive Models

Overview

Previous books provided the background on the learning of local descriptive models from unlabeled (or single-label) high-dimensional data. The relevance of guaranteeing the flexibility and robustness (*Books III-IV*), as well as the statistical significance (*Book V*), of these models was largely discussed and motivated for biomedical and social domains. When moving from these data contexts towards labeled data contexts, the learning of (class-conditional) descriptive models can be seen as a direct byproduct of previous contributions and, therefore, it is typically less relevant. Instead, the learning of decision models becomes the primary goal.

This book targets the specific task of learning effective (associative) classification models from both tabular and structured data contexts. Illustrative biomedical decisions include the discrimination of biological phenotypes, the anticipation of medical conditions, and the learning of biological and clinical markers. Illustrative social decisions include the classification of individuals' behavior and profile, the evaluation of (web) contents, and the support for trading, administrative and commercial decisions. These applications typically rely on high-dimensional data, where the number of features (such as genes, clinical features, contents or actions) may exceed the number of class-conditional observations (such as samples and individuals).

In this context, the learning should be able to minimize the susceptibility of the learning function to: 1) overfit the input data by guaranteeing that uninformative regions are discarded, and 2) underfit the input data by guaranteeing that the decisions are made from statistically significant regions. This observation, together with the gathered evidence of the relevance of learning from local regions and the limited role of dimensionality reduction and sparse kernels, stress the importance of learning local classification models from high-dimensional data contexts.

As such, classification, the task of learning a mapping model to label unlabeled observations from a training set of labeled observations, becomes centered on informative and discriminative regions. In this learning context, the mapping model is referred as an associative model and the learning task (associative classification) is driven by three major requirements:

- effective discovery of relevant regions, where the relevance is essentially related with their homogeneity (coherency and quality), discriminative power and statistical significance;
- adequate scoring and composition of regions (training function);
- robust matching and scoring schema to test a new observation against the learned model (testing function);

The first part of this book (*Chapters 1-4*) addresses these requirements to guarantee an adequate learning from both tabular data **[R5.1]** and structured data **[R5.2]**.

However, in order to address the hypothesis of this thesis, we need to guarantee not only the accuracy, but also the statistical significance of classification decisions **[R5.3]**. In other words, the focus should not be uniquely placed on the optimization of the average performance of classifiers, but also on the minimization of the performance variability. Guaranteeing the statistical significance of classification decisions is of increasing importance to validate biomarkers and computer-aided decisions associated with medical decisions, as well as to support trading, marketing and other social initiatives with potential high costs. *Chapters 5-6* address this additional requirement, thus minimizing the propensity of associative classifiers to underfit high-dimensional data (inference of decisions from non-significant regions and/or from a subset of all relevant regions).

Under these contributions, we guarantee the learning of robust (associative) classifiers with controlled risks of under/overfitting data. However, since the majority of real-world decisions change over time, it is increasingly

important to temporally frame decisions to solve predictive tasks **[R5.4]**. In this context, *Chapter 7* extends previous contributions to answer the task of classifying an attribute of interest across different time periods, referred as multi-period classification.

Follows a brief discussion of the problems tackled by each chapter of this book.

[R5.1] *Chapters 1-3* guarantee an adequate learning of associative classifiers from (high-dimensional) *tabular data*. *Chapter 1* addresses the major criticisms of existing associative classifiers, including: scarcity of matchings, inability to adequately score noisy regions, inadequate scoring in the presence of imbalanced data, biases towards small (non-significant) regions, inappropriate space exploration, absence of adequate dissimilarity guarantees between regions, and inability to model regions discriminating more than a single class. For this aim, we augment the pattern-based biclustering contributions proposed in *Book III* with adequate discriminative criteria for the selection of relevant regions. These regions are composed within an associative model using new integrative scores and matched against testing observations using essential relaxations.

Chapter 2 extends these contributions when the learning is driven by regions with varying properties, including flexible coherency assumptions, coherency strength and quality. Discovery, training and testing functions are adequately revised for this end, and their relevance when learning from biological and social datasets assessed.

Chapter 3 explores advanced aspects of associative classification from tabular data. First, we propose associative classifiers able to learn from sparse data. Second, we provide principles to learn ensemble models when the input data is characterized by the presence of local and global regularities. Third, we discuss the benefits and limitations of learning classifiers from stochastic (versus deterministic) biclustering models. Fourth, we show the applicability of the classifiers from this book towards tabular data with non-identically distributed features. Finally, we extend the previously proposed classifiers to effectively accommodate background knowledge.

[R5.2] *Chapter 4* extends the classification scope towards (high-dimensional) *structured data*. In this context, we aim to develop effective classifiers to learn from labeled observations from a data space with an arbitrary-high multiplicity of temporal attributes, able to model regions with discriminative, temporal and integrative (cross-attribute) regularities. To adequately learn from these discriminative regions, both deterministic and stochastic classifiers are proposed and their behavior confronted. We further specialize these learning functions towards data contexts given by three-way time series and multi-sets of events, where these regions are respectively associated with discriminative cascades and arrangements of informative events.

[R5.3] *Chapters 5-6* extend previous contributions in order to guarantee the statistical significance of classification decisions. The application of associative classifiers over high-dimensional data often leads to decisions inferred from small regions, which are typically informative or discriminative by chance. In this context, three additional requirements need to be tackled:

- guarantee the statistical significance of discriminative regions **[R5.3.1]**;
- assess the impact of associative training and testing functions on the significance **[R5.3.2]**;
- optimize classification performance while providing guarantees of significance **[R5.3.3]**.

Chapter 5 measures the impact of learning from regions with varying statistical significance and revises the learning functions accordingly. For this end, since the contributions proposed throughout *Book V* are insufficient towards this end, they are extended to guarantee not only that the probability of a region to occur deviates from expectations (significantly informative), but also that its support significantly varies between class-conditional data partitions (significantly discriminative).

Despite the relevance of these contributions, they are insufficient to guarantee statistically significant decisions. Illustrating, the commonly applied pruning procedures during the training stage and matching relaxations during

the testing stage often interfere with the guarantees of significance. In this context, *Chapter 6* measures the impact of these learning options on the propensity towards false positive and negative decisions per class and propose principles to minimize them. Furthermore, it combines both accuracy (average error) and significance (variability of error) views since the blind optimization of significance levels can impact accuracy. Finally, this chapter conducts a systematic experimental assessment of the benefits and limitations of the enhanced classifiers for different high-dimensional data domains, providing supporting evidence for the validation of the thesis hypothesis.

[R5.4] The classifiers enhanced throughout this book place a single decision per testing observation. However, real-world decisions change over time. As such, *Chapter 7* provides principles to guarantee that classifiers are able to learn from structured codomains given by sequences of classes. Despite the relevance of this task (referred as multi-period classification) to answer a wide-set of real-world predictive problems, existing research fails to model the stochastic dependencies between the periods under classification and requires dedicated learning functions (preventing the use of the previously proposed classifiers). In this context, we provide a formal view on this task and propose new methods able to surpass the limitations of peer attempts driven from long-term prediction and multi-label classification.

Index of Requirements and Contributions

Tables 2-6 exhaustively list the proposed contributions throughout this book.

Table 1: Major contributions to learn effective associative classifiers from high-dimensional tabular data (*Chapter* VI-1).

R5: Learning effective (associative) classifiers for high-dimensional data;
R5.1: Learning effective associative classifiers from tabular data contexts;
- **C5.1a**: Structured view on the limitations and potentialities of associative classification;
- **C5.1b**: Systemic analysis of the impact of varying coherency assumption, coherency strength, quality, discriminative power and significance on the performance of associative classifiers;

R5.1.1: Effective discovery and selection of relevant regions;
- **C5.1.1a**: New weighted notion of support to adequately assess the discriminative power of noisy regions;
- **C5.1.1b**: Efficient discovery of regions able to discriminate groups of classes and generation of rules with disjunctions of labels;
- **C5.1.1c**: New discriminative criteria able to deal with data imbalance and rules with disjunctions of labels;
- **C5.1.2a**: Adequate exploration of high-dimensional data spaces: focus on diverse sets of regions with dissimilarity guarantees;
- **C5.1.2b**: Revised learning methods that effectively and efficiently use the discriminative power to guide the space exploration;
- **C5.1.2c**: Integration of discriminative power, homogeneity and significance views to guide the learning;

R5.1.2: Effective scoring and composition of regions (training);
- **C5.1.2.1**: New integrative training scores able to effectively combine the discriminative power, size and quality of a region;
- **C5.1.2.2**: Adequate data structures relating regions according to their properties and scores for an efficient testing;

R5.1.3: Effective matching and labeling of new observations against the structure of scored regions (testing);
- **C5.1.3.1**: Relaxations on the matching criterion to guarantee an adequate number of matches per testing observations;
- **C5.1.3.2**: Effective calculus to compute class strength, able to deal with: 1) matched rules with disjunctions of labels, and 2) aligned with the proposed integrative scores;

R5.1.4: Learning classifiers from regions with varying homogeneity;
R5.1.5: Effective learning functions to classify sparse data;
R5.1.6: Effective learning from both global and local regularities underlying data;
R5.1.7: Understand the impact of learning classifiers from stochastic local descriptive models;
R5.1.8: Adequate learning of classifiers in the presence of background knowledge;
R5.2: Learning effective associative classifiers from structured data contexts;
R5.3: Significant classification decisions from high-dimensional data;
R5.4: Multi-period classification: extend previous contributions for the learning of sequences of classes;

Table 2: Proposed contributions on the learning of classifiers from regions with varying homogeneity (*Chapter* VI-2).

R5.1.4: Learning classifiers from regions with varying homogeneity;
R5.1.4.1: Learning classifiers from regions with varying coherency;
- **C5.1.4.2.1a**: Penalization schema for non-constant regions based on their degree of flexibility (to guarantee the diversity of regions and tackle domination of learning by a subset of relevant regions);
- **C5.1.4.2.1b**: Integrative discovery of discriminative regions with multiple coherency assumptions (lift and statistical views);
- **C5.1.4.2.2a**: Extended testing score to compute the strength of each class from multiple interestingness criteria;
- **C5.1.4.2.2b**: New matching criteria to test observations against non-constant regions based on the allowed adjustment factors;

R5.1.4.2: Learning classifiers from regions with varying quality;
- **C5.1.4.2a**: Extended discovery guarantee the presence of (discriminative) regions with varying quality and coherency strength;
- **C5.1.4.2b**: Revised integrative score to better weight the quality of a region (deviations from pattern expectations);

Table 3: Contributions to learn classifiers from sparse data, data with regularities of varying extent, stochastic descriptors of data, complex tabular data and data domains with available background knowledge (*Chapter* VI-3).

R5.1.5: Effective learning functions to classify sparse data;
- **C5.1.5a**: Principles to bypass the interpretation of missing elements from structurally sparse data (surpassing the need for imputation);
- **C5.1.5b**: Sound classification from regions with arbitrary-high number of true and/or false missings;
- **C5.1.5c**: Extension (and shown compliance) of training and testing functions to correctly handle sparse data;
- **C5.1.5d**: Adequate data structures and searches for an efficient classification of sparse data;

R5.1.6: Effective learning from both global and local regularities underlying data;
- **C5.1.6**: Extended associative classifiers with new voting schema to combine the (probabilistic) decision outputs of global kernels;

R5.1.6.1: Minimize bias from decisions with low confidence;
- **C5.1.6.1a**: Exclusion of class-conditional observations without clear local regularities from associative learning, and exclusion of class-conditional observations without clear global regularities from global kernels;
- **C5.1.6.1b**: Exclusion of decisions with low confidence (e.g. few matches) and loose strength (e.g. contradictory matches) from voting;

R5.1.7: Understand the impact of learning classifiers from stochastic local descriptive models;
- **C5.1.7a**: Principles on how to use membership vectors to guarantee an adequate and easily parameterizable coverage of the data space;
- **C5.1.7b**: Principles to use membership vectors to guarantee more accurate scores and prevent the scarcity of matched regions;
- **C5.1.7c**: Preliminary learning functions to infer decisions from class-conditional parametric models;

R5.1.8: Adequate learning of classifiers in the presence of background knowledge;
- **C5.1.8.1a**: Principles for guiding classifiers in the presence of annotations extracted from knowledge bases and literature;
- **C5.1.8.1b**: Demonstrated compliance of FleBiC to effectively incorporation of constraints with nice properties (succinct, monotonic, anti-monotonic, convertible and prefix-monotone) targeting the informative data regularities;
- **C5.1.8.1c**: Incorporation of a new class of constraints with nice properties targeting discriminative and class-conditional regularities;

Table 4: Proposed contributions to learn associative classifiers from structured data (*Chapter* VI-4).

R5.2: Learning effective associative classifiers from structured data contexts;
- **C5.2a** Survey on pattern-based and stochastic learning from temporal data and integrative strategies to handle multiple attributes;
- **C5.2b** Comparison of deterministic and stochastic learning functions and principles for their adequate selection;

R5.2.1: Applicability to varying data structures;
- **C5.2.1.1a**: Definition of an integrative and temporal data structure conducive to learning;
- **C5.2.1.1b**: Principles to map multi-dimensional/relational data, multi-sets of events and three-way time series into the target structure;
- **C5.2.1.1c**: Extension of mappings for labeled data contexts and principles for the retrieval of annotations;
- **C5.2.1.2**: Principles to handle data structures characterized by a mixture of non-identically distributed temporal and static attributes;

R5.2.2: Effective pattern-based classifiers;
- **C5.2.2**: Extension of contributions *C3.1-2* to learn associative classifiers from discriminative, integrative and temporal patterns;

R5.2.2.1: Adequate discovery of informative and discriminative regions;
- **C5.2.2.1a**: Revised support notion to guarantee its tolerance to structural and temporal misalignments;
- **C5.2.2.1b**: Extended discovery driven by discriminative criteria based on variant of rule's lift (based on the revised support);
- **C5.2.2.1c**: Efficient composition of rules with disjunctions of labels in the consequent;

R5.2.2.2: Adequate scoring and composition of regions (training);
- **C5.2.2.2.1**: New integrative score based on the (revised) support, length and lift of the target integrative and temporal regions;
- **C5.2.2.2.2a**: Composition of regions within a tree structure promoting an adequate traversal for an efficient detection of matches;
- **C5.2.2.2.2b**: Pruning of regions to deal with heightened imbalance on the score and/or number of regions per class;

R5.2.2.3: Effective testing of new observations;
- **C5.2.2.3a**: New matching criteria sensitive to both structural and temporal misalignments;
- **C5.2.2.3b**: Degree of matching relaxations dependent on the number, score and class-consistency of matched regions;
- **C5.2.2.3c**: Penalization factor based on the temporal mismatches between a testing observation and the learned regions;
- **C5.2.2.3d**: New class strength calculus based on the proposed integrative score;

R5.2.2.4: Adequacy of behavior for regions given by discriminative responses;
- **C5.2.2.4**: Specialization of the proposed behavior to adequately model cascades and arrangements of events, including: a) postprocessing procedure, b) module aggregation and causality identification, and c) principles to foster efficiency;

R5.2.2.5: Advanced weighting of occurrences along time (selective/decaying memory);
- **C5.2.2.5**: Easily parameterized behavior to prioritize occurrences on certain time periods (according to linear or exponential functions) to attenuate the impact of older discriminative events on decisions;

R5.2.3: Effective stochastic classifiers;

R5.2.3.1: Modeling regions of interest associated with temporal and integrative views;
- **C5.2.3.1a**: Reuse of contributions *C3.3* for the class-conditional learning of generative models sensitive to local regularities;
- **C5.2.3.1b**: Extension of contributions towards classification by either: 1) decoding of regions for classic associative classification, or 2) testing the likelihood of new observations to be described by the learned class-conditional models (default);
- **C5.2.3.1c**: Principles for an efficient testing (based on pruning and non-redundant computation of the sum of joint probabilities);

R5.2.3.2: Adequacy of behavior when learning from multi-sets of events and three-way time series;
- **C5.2.3.2a**: Customized behavior for three-way time series to support: 1) incremental learning, 2) modeling of numeric data, and 3) correct interpretation of multi-item assignments;
- **C5.2.3.2b**: Specialization for multi-sets of events: proper initialization of delimiter emissions and efficient decoding of time frames;

Table 5: Contributions for the learning of classifiers from significantly discriminative and informative regions and with minimized propensity to overfit and underfit high-dimensional data (*Chapters* VI-5 and VI-6).

R5.3: Significant classification decisions from high-dimensional data;
R5.3.1: Effective classification based on statistically significant regions;
C5.3.1.1: Structured view on the major limitations of state-of-the-art work towards this end;
C5.3.1.2a: Statistical view of the discriminative power of regions from tabular and structured data;
C5.3.1.2b: Integrative statistical views on the significance of the informative and discriminative power of a region;
C5.3.1.3: Principles to guarantee adequate learning for data with a scarcity of simultaneously informative and discriminative regions;
C5.3.1.4a: Revised associative classifiers from tabular data: revised region discovery and scoring schema to accommodate significance criteria based on C4 contributions;
C5.3.1.4b: Discussion of whether alternative stochastic learners are able to similarly provide guarantees of significance;
C5.3.1.5: Revised associative classifiers from structured data based on C4.5 contributions;
C5.3.1.6a: New class of decision trees based on revised building of trees with paths given by significant regions whenever possible;
C5.3.1.6b: Revised random forests to further address the underfitting propensity of local classifiers;
R5.3.2: Training and testing functions with guarantees of statistical significance;
C5.3.2.1a: Principles to assess the impact of associative training functions on the number of false positives and negatives;
C5.3.2.1b: Principles to assess the impact of testing functions (matching criteria) on the number of false positives and negatives;
C5.3.2.1c: Revised behavior of associative classifiers based on the proposed principles;
C5.3.2.2: Extensibility of these contributions for alternative classifiers;
R5.3.3: Indicators of the statistical significance guarantees of classification decisions to support real-world decisions;
C5.3.3.1: Annotation of rules with an integrative statistical measure of their discriminative and informative significance;
C5.3.3.2: Annotation of classification decisions with an indicative score of their significance based on the statistical guarantees provided by the discovery, training and testing functions;
R5.3.4: Integrative view of accuracy (average error) and significance (variability of error);
C5.3.4.1a: Principles to jointly optimize accuracy and significance views;
C5.3.4.1b: Principles to learn from data with few significant regions and with imbalanced number of rules per class;
C5.3.4.2: Extensive experimental evaluation (using C1 methodology) of the proposed learning methods (using C2-C5 contributions) over high-dimensional data from distinct domains;

Table 6: Contributions to classify a class at different time periods for predictive tasks (*Chapter* VI-7).

R5.4: Multi-period classification: extension of previous contributions for the learning of sequences of classes;
R5.4.1: Task formalization and evaluation (standardize and validate upcoming contributions);
C5.4.1.1a: Data-indepent and model-independent formalization of the multi-period classification task;
C5.4.1.2a: Evaluation metrics for assessing multi-period classifiers from extended three-dimensional confusion matrices;
C5.4.1.2b: Extended assessment for sequences of ordinal labels;
C5.4.1.2c: Distance metrics to account for temporal misalignements and error accumulation between estimated and observed sequences;
R5.4.2a: Modeling the stochastic dependencies underlying the sequence under classification;
R5.4.2b: Embedding existing (single-label) classifiers able to learn from tabular/structured data;
C5.4.2.1: New hybrid method able to trade-of the properties of iterative and direct single-output methods from long-term predictions;
C5.4.2.2a: Cluster-based multi-period classifier (adequate reduction and recovery of the space of sequences) able to model dependencies between classes and guarantee indepndence from the underlying single-label classifier;
C5.4.2.2b: Dynamically parameterizable behavior of the cluster-based methods (based on the number of periods, labels, observations, and diversity and representativity of sequential behavior) to minimize the number of false positive and false negatives;
C5.4.2.3a: Variant based on the segmentation of the sequence of classes to minimize the flexibility issues of multiple-output peers;
C5.4.2.3b: Principles for segmenting sequences based on sensitivity analysis, stochastic properties of the observed sequences, the periodicities and local stationarity when available, and the analysis of clustering error (within-cluster sum of squares);
C5.4.2.3c: Variant combining a moving sliding-windows with voting schema to guarantee a more accurate modeling of the true stochastic dependencies between the periods under prediction.

1

Effective Associative Classification from Discriminative Biclustering Models

High-dimensional biomedical and social data are characterized by the presence of local regularities, whose discovery has been largely motivated throughout the previous books. An additional common property of these data contexts is the presence of a large extent of either uninformative or non-discriminative regions. In this context, learning classification models from high-dimensional data is challenged by the need to focus the learning in regions of interest. This chapter aims to address this challenge in the context of high-dimensional tabular data, where regions can be flexibly described by biclusters (subsets of features and observations) with specific homogeneity, significance and discriminative power.

Different strategies have been considered in literature to reduce the complexity of the learning function and its propensity to overfit or underfit high-dimensional data, including feature selection, dimensionality reduction transformations and sparse priors. Despite their relevance, these options are neither able to flexibly select relevant regions nor discard non-relevant regions [512, 660]. To address this problem, classifiers able to approximate local regularities, such as associative classifiers, can be considered to focus the learning on specific regions of interest given by discriminative biclusters [99, 497, 634].

However, and despite the large extent of research available, associative classification still suffers from three major drawbacks. First, existing associative classifiers are not able to adequately learn from noisy and real-valued regions. This is related with the fact that the contributions on this field were developed in the context of transactional and discrete data contexts [99, 497], where discriminative patterns were seen as the basis for the learning. Understandably, regions may show higher support or discriminative power at the cost of tolerating noisy elements. In this context, it is also important to guarantee that these noisy regions do not jeopardize the learning.

Second, in the presence of labeled tabular data with more than two classes, regions may not be able to discriminate a single class, yet able to discriminate a subset of overall classes. Illustrating, a given pattern may be inhibited on a specific class or elicited on a subset of all classes. Although regions with such pattern are not able to discriminate a single class, they may have an important value for learning.

Third, one of the criticisms of associative classification is related with the fact that new (testing) observations may not be well-described by the selected discriminative regions from labeled (training) observations. This often leads to decisions with low level of confidence due to both the lack of evidence or class-contradictory evidence.

Finally, existing associative classifiers are not able to adequately explore the data space (learning is easily jeopardized by the presence of a few large regions) and provide dissimilarity guarantees [99, 131].

This chapter aims to provide a structured view on the existing efforts towards associative classification from discriminative biclusters, and further extend these efforts with new principles to address the identified challenges. This is done in line with three major requirements associated with the learning of associative classifiers: 1) effective discovery and selection of regions of interest, 2) adequate training functions for their scoring and composition, and 3) robust testing functions to guarantee an accurate labeling of unlabeled observations. The major contributions of this work are seven-fold:

- structured view on the contributions and criticisms of existing associative classification models;

- new weighted notion of support and lift to adequately assess the discriminative power of noisy regions;
- new discriminative criteria able to deal with data imbalance and rules with disjunctions of labels;
- discovery of dissimilar regions with varying coherency, quality and significance;
- efficient composition of rules based on regions able to discriminate groups of classes;
- new integrative training scores able to effectively combine the discriminative power of a region and its additional properties;
- more effective calculus to compute class strength during the testing stage, able to deal with disjunctions of labels and the proposed integrative scores;
- effective relaxations to guarantee an adequate number of matches per testing observations.

These contributions are integrated within a new associative classifier, BiC (Biclustering-based Classifier). We provide initial empirical evidence of the relevance of the listed contributions.

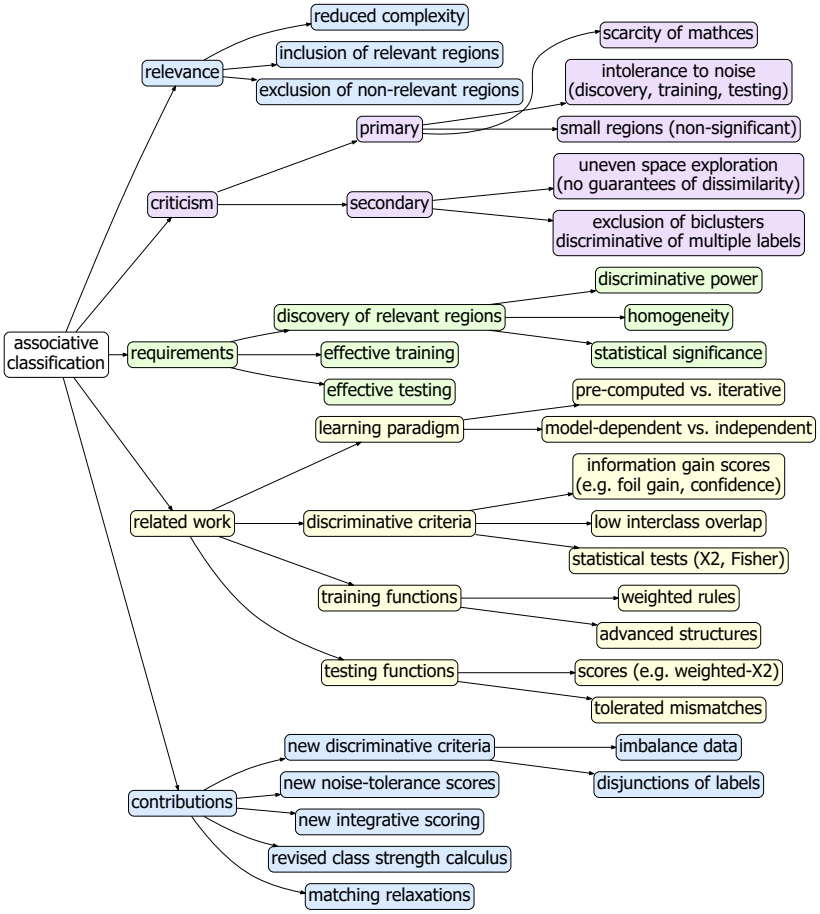


Figure 1.1: Requirements and principles for learning associative classifiers from high-dimensional spaces.

Figure 1.1 structures the tackled challenges and contributions for the adequate learning of associative classifiers from high-dimensional data. Accordingly, this chapter is organized as follows. *Section 1.1* provides the background on the target problem. *Section 1.2* surveys the contributions and limitations from state-of-the-art research on associative classification. *Section 1.3* describes the solution space by proposing BiC. *Section 1.4* gathers experimental results that support the utility of BiC. Finally, concluding remarks and implications are synthesized.

1.1 Background

This section motivates and formulates the task of learning effective associative classifiers from flexible biclustering models.

A *labeled tabular dataset* $\mathbf{A}$ is described by n labeled observations (rows) $\mathbf{a}_i=(\mathbf{x}_i \in \mathcal{X}, c_i \in C)$, where $\mathcal{X}=\{\mathcal{Y}_1,..,\mathcal{Y}_m\}$ is a space of m features (columns) and C is a finite set of either nominal or ordinal classes. Given $\mathbf{A}$, consider the set of n observations to be $\mathbf{X}=\{\mathbf{x}_1,..,\mathbf{x}_n\}$, and the set of m feature-vectors to be $\mathbf{Y}=\{\mathbf{y}_1,..,\mathbf{y}_m\}$. In this context, $\mathbf{A}$ is defined by $n \times m$ elements, $a_{ij} \in \mathcal{Y}_j$, corresponding to the observed value for the $\mathbf{x}_i$ observation and $\mathbf{y}_j$ feature.

As previously defined, a region from a tabular data space $\mathbf{A}$ is a bicluster, $\mathbf{B}=(\mathbf{I},\mathbf{J})$, i.e. a subspace given by a subset of observations, $\mathbf{I} \subseteq \mathbf{X}$, and features, $\mathbf{J} \subseteq \mathbf{Y}$. A region of interest is given by a bicluster satisfying specific criteria of *homogeneity* (coherence and quality), statistical *significance* and *discriminative power*.

1.1.1 Problems of Classification from High-Dimensional Data

As illustrated in Figure 1.2, learning from a high-dimensional data space is complex since many of its elements are either non-informative or non-discriminative [512, 660]. In this context, different contributions have been proposed to guarantee an adequate learning focus.

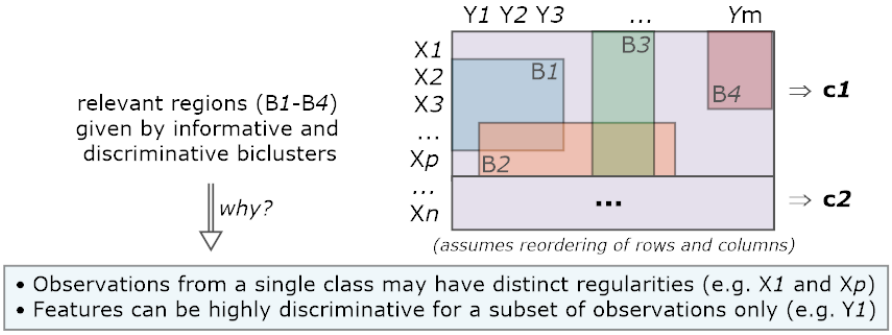


Figure 1.2: Learning from high-dimensional data spaces: importance of discovering relevant regions given by coherent, statistically significant and discriminative biclusters.

First, feature selection focus on one region given by a subset of features only, $\mathbf{B}=(\mathbf{X},\mathbf{J})$ (where $\mathbf{J} \subseteq \mathbf{Y}$). Understandably, its use is insufficient to address the target problem since regions that are highly discriminative on a compact subset of overall observations are prone to be excluded. Due to the inherent complexity of class-conditional regularities in real-world data contexts, the selection of these regions is critical. Furthermore, feature selection is commonly applied in the absence of statistical significance criteria. As such, a compact subset of features can be selected as long as the combination of their values discriminate a particular (or all) classes, yet the observed combination of values may be discriminative by chance. In *Section I-1.2*, an instantiation of this last problem was provided, as well as a closer look on the challenges of applying feature selection either as a filter or as a wrapper.

Second, an alternative way of reducing dimensionality is to approximate and apply a mapping reduction function, also referred as a projection, from the observed data space into a new data space with lower dimensionality. Although these functions were originally proposed to transform real-valued data spaces, $\phi : \mathbb{R}^m \rightarrow \mathbb{R}^d$ where $d<m$, additional attempts (developed in the context of hyper-dimensional mappings and support vectors) have been proposed for their application over data spaces with non-identically distributed features [71]. However, even in the presence of non-trivial mapping functions, to our knowledge there are not projections able to select multiple relevant regions and flexibly neglect non-relevant regions from the input data space.

Finally, stochastic learners have been augmented with sparse priors to improve their ability to learn from high-dimensional data contexts. In these data contexts, irrelevant and redundant parameters rapidly converge to zero with these priors [378, 214]. *Basics I-1.17* and *Pointer I-1.18* provide respectively a description of the properties of these learners and a compact overview of some of the state-of-the-art options. However, these classifiers suffer from a similar problem as previous alternatives. Since sparsity is determined by the parametric model, they cannot flexibly select regions of interest. As such, the sparsity is primarily used to either discard non-relevant features and/or specific ranges of values per feature.

In the presence of high-dimensional data, the majority of existing classifiers are applied with on one or more of the previously enumerated options. In this context, these classifiers are insufficient to guarantee an effective learning since they are associated with: *1)* the inclusion of non-relevant regions (promoting *overfitting*[1]), and *2)* the exclusion of relevant regions (promoting *overfitting*[2]) [512, 105].

1.1.2 Motivating Associative Classification

As largely motivated in *Section I-1.2*, biological and social data is characterized by the presence of flexible structures of regions. Learning classifies from regions given by (discriminative) biclusters is increasingly relevant to learn biological and clinical markers [429, 122], classify users' behavior [257], evaluate contents [159] and support trading and commercial decisions [334, 35]. These observations stresses the need to move from global towards associative models.

Def. 1.1 Given a set of observations in $\mathcal{X}$ labeled with a class in C, a *descriptive model* (either locally or globally) approximates the class-conditional regularities of the space, $M(\mathcal{X}|C)$. In particular, an **associative model**, also referred as a *discriminative biclustering model*, is a composition of p association rules, $\{r_1, .., r_p\}$, where each rule $r_i : \mathbf{B}_i \Rightarrow C_i$ maps a region of interest $\mathbf{B}_i$ (rule's antecedent given by a discriminative bicluster) with a set of labels $C_i \subset C$ (rule's consequent).

In this context, an associative model is associated with a structure of rules given by regions of interest characterized by the coherence (type of homogeneity), quality (homogeneity strength), significance, discriminative power of the underlying biclusters. The *structure* of an associative model is essentially defined by the number, size and positioning of the biclusters [429]. The *coherence* of a bicluster is defined by the observed correlation of values. The *quality* of a bicluster is defined by the type and amount of accommodated noise. The statistical *significance* of a bicluster determines its deviation from expectations. The *discriminative power* of a bicluster measures its probability to only occur significantly for a subset of overall classes. In this context, an increase in the coherence strength (respecting Def.II-3.1 and Def.II-3.2), quality (low η_{ij} according to Def.II-3.2), statistical significance (according to *Book V*) and discriminative power $P(C_k|B_k)$ of a given bicluster denotes an increase in the relevance (and thus an increase score) of the associated rule.

In the presence of a set of discriminative biclusters, moving from descriptive to classification settings becomes a matter of defining effective training and testing functions.

Def. 1.2 A *classification model* is a mapping function between observations and classes, $M : \mathcal{X} \rightarrow \Sigma$, to label (unlabeled) observations. An **associative classification model** defines a matching criteria M to label observations against a (possibly pre-computed) *associative model*.

An illustrative simplistic *associative classifier* is one that learns rules associated with biclusters above certain support (number of observations) and confidence (discriminative power), and that classifies a new/testing observation based on the sum of class-conditional scores associated with the biclusters that best describe its values. *Basics I-1.12* provides an illustrative instantiation of a similar associative classifier.

1.1.3 Limitations of Existing Associative Classifiers

Despite the argued need for associative classification to learn from high-dimensional data, some criticisms have been pointed in the past with regards to their performance. This is a result of three major observations. First, their inability to focus on statistically significant regions of interest. The majority of existing associative classifiers are primarily concern with the discriminative power of the modeled regions, and therefore do not guarantee whether or not these regions occur by chance. Empirical evidence from the application of decision trees and pattern-based classifiers over high-dimensional datasets are associated with decisions made from regions with 3 to 5 features, as

[1] Also referred as propensity to underfit relevant regions (≡overfit input data).
[2] Also referred as propensity to overfit the incomplete set of regions (≡underfit input data).

the combination of values on such features become able to discriminate a given class. Understandably, such small regions are not able to provide guarantees of statistical significance and therefore the resulting classifiers show undesirably high risk of underfitting towards the input dataset, which penalizes decisions.

Related with this problem, there is the fact that associative models have its roots on transactional data analysis with discriminative pattern mining [409, 184], and therefore are not natively prepared to deliver regions tolerant to noise. Although extensions have been proposed to tackle this observation, these extensions are not able to weight the relevance of regions by their quality. As a result, larger regions due to the accommodation of noise tend to score high due to their size and (possibly) discriminative power, even if that comes at a cost of quality. In this context, these regions can unfairly jeopardize the learning.

Second, the performance of associative classifiers is highly hampered by the inability to guarantee that a testing observation has a reasonable number of matchings in order to determine its label with confidence. This is a result of two major issues: 1) inadequate space exploration where many informative and discriminative regions of the input data space are discarded; and 2) restrictive matching criteria. In this context, empirical evidence shows that testing observations are often associated with either a few (possibly contradictory) matchings or no matchings at all.

Finally, the majority of existing associative classifiers rely on discriminative patterns without guarantees of dissimilarity. In this context, the existence of similar rules can jeopardize matching scores and reinforce an incorrect labeling decision.

An additional important observation, although with less impact on the gathered criticisms, is the fact that the decision rules learned from existing associative classifiers are only able to discriminate a single class. However, in the presence of labeled tabular data with more than two classes, regions may not be able to discriminate a single class, yet be able to discriminate a subset of overall classes with high confidence.

1.2 Related Work

The discovery of discriminative regions has been mainly driven by research on discriminative pattern mining [99] and, more recently, by research on discriminative biclustering [122]. Below, we cover the contributions according to the introduced requirements: discovery of discriminative regions (*Section 1.2.1*), their composition (*Sections 1.2.2*) and use to label observations (*Sections 1.2.3*).

1.2.1 Discovery of Discriminative Regions

Discriminative Pattern Mining. The discovery of discriminative patterns has received a wide-attention in recent years, with multiple works providing categorizations on how to find discriminative patterns and on how to learn classification models from these regions [99, 660, 497]. Bringmann et. al [99] categorize associative classifiers along two axes: whether they learn from a pre-computed set of regions or iteratively discover new (or extend existing) regions, and whether they select regions independently or guided by the properties of the target learning function.

On the first axis, earlier studies focus on mining all frequent patterns (constant biclusters) per class at a time in order to compose rules with high discriminative power. The input data is thus partitioned and the mining task applied independently on each partition. From the computed set of patterns, many metrics have been proposed to fix adequate class-conditional support levels and to assess the correlation strength ϕ between a pattern and a class, as well as to relate both these views within a single score [99]. Illustrative associative classifiers with alternative scoring schema include CBA [409] (ϕ=confidence), classifiers based on emerging patterns [184] (ϕ=growth), CMAR [401] (ϕ=χ^2), CPAR [696] (ϕ=foil gain) and RCBT [146] (ϕ based on top-k covering rule groups) and lazy classifiers that only retrieve classification rules once a test instance is given [649]. Alternative correlation scores can be given by information gain, Fisher score and lift [99].

However, even in the presence of constraints and condensed pattern representations to deliver compact sets of distinct patterns, these methods easily became computationally expensive for large or dense data spaces with low support thresholds. An alternative to avoid the generation of the exhaustive pattern set is to perform branch-and-bound or iterative-deepening searches [497]. These searches extend the (anti-)monotonic principles to guarantee that only regions with discriminative are explored. Examples of associative classifiers parameterized with these searches are Harmony [656], DDPMine [131], MbT [206], and decision trees.

On the second axis, model-dependent approaches rely on the knowledge regarding the behavior and expected outputs of a given classifier to affect the discovery and selection of patterns [99]. As such, their behavior contrasts with model-independent approaches, which are not able to offer ground guarantees on whether the discovered set of patterns is the most adequate.

Recently, classifiers appearing in both sides of these axes have been proposed using diverse sets of patterns by relying on ensemble models and enriched with machine learning-inspired sampling and verification techniques [715].

Despite the listed contributions, discriminative pattern mining suffers from three major limitations: 1) inability to discover real-valued regions none prone to discretization problems; 2) inability to model non-constant regions; and 3) inability to discover regions with arbitrary levels of noise.

Discriminative Biclustering. The previous scope of research can be enlarged by considering flexible and noise-tolerant regions through discriminative biclustering. Many biclustering algorithms have been proposed to find biclusters with varying structure, coherence and quality [429]. In Table III-1.3 we surveyed varying methods with regards to their optimality guarantees, search paradigm and the properties of the underlying merit function (the means to guide the search towards the discovery of biclusters with certain desirable properties of interest).

Biclustering can be extended for associative classification by defining $|C|$ class-conditional searches and adequately scoring the discriminative power of biclusters. Recently, Odibat and Reddy [497] proposed principles to push the discriminative criteria deep into the biclustering task in order to narrow the search space.

Discriminative biclustering methods have been recently proposed with alternative score variants, including FDCluster [661], DRCluster [660], among others [122, 612]. Di-RAPOCC [497] considers a bicluster to be discriminative if it has high confidence and low inter-class overlapping (biclusters discovered in one class should have a minimal number of rows in other classes).

The major problems with the existing discriminative biclustering approaches are three-fold. First, restrictions on the allowed number, size and positioning of biclusters [310], degrading associative classification in high-dimensional data contexts. An exhaustive coverage of (potentially relevant) regions with varying properties of interest is critical to guarantee that testing observations match a substantial number of regions (thus increasing the confidence of decisions). Second, the few (discriminative) biclustering algorithms able to discover regions with parameterizable quality and coherency show critical limitations (listed in Table III-6.1). Finally, the existing algorithms do not guarantee the statistical significance of the discovered discriminative biclusters.

In this context, although the biclustering algorithms proposed throughout *Book III* (and enriched with statiscal guarantees of significance in *Book V*) can be used to surpass these limitations, they are neither natively prepared to incorporate discriminative criteria when observations are labeled nor to extend the proposed statistical tests to guarantee the significance of a discriminative rule.

1.2.2 Associative Training Functions

Given a set of regions of interest, different composition functions have been considered to train a classification model, ranging from simplistic sets of weighted association rules of the type $\mathbf{B} \Rightarrow c$ to more structured models. Examples include the integration of these regions within naive Bayesian classifiers [396] and decision trees [248], as well as support vector machines (SVMs) over feature-spaces given by multi-class discriminative subspaces

[612]. The work by Carreiro et al. [122] studies alternative ways to compose discriminative biclusters to perform classification.

Although many training variants exist, let us consider the training steps of CMAR [401] as the prototypical case. First, association rules are learned from regions satisfying minimum support (*sup*) and confidence (*conf*) thresholds. Second, each rule is scored using a χ^2 test and inserted in a tree structure (CMAR-tree) according to their priority. A rule $R_1 : \mathbf{B}_1 \Rightarrow c$ is said to have priority over $R_2 : \mathbf{B}_2 \Rightarrow c$ if $R_1 \supseteq R_2$or if $conf(R_1) > conf(R_2) \vee (conf(R_1){=}conf(R_2) \wedge sup(R_1) > sup(R_2)) \vee (conf(R_1){=}conf(R_2) \wedge sup(R_1){=}sup(R_2) \wedge |\mathbf{I}_1| < |\mathbf{I}_2|)$. Finally, the CMAR-tree is pruned based on the observed priority of rules to guarantee a balanced number of rules across classes.

In this context, adequate scoring methods are required to weight the interestingness of each association rule, ranging from simple metrics, such as the support-and-length (an indicator of the subspace's significance) and the confidence (an indicator of discriminative power) of each rule. More complex scoring methods include relevance criteria derived from probabilist induction [634] and optimization metrics based on confusion matrices [204].

Recently, extended association rules have been proposed to allow for both disjunctive patterns and disjunctions of classes per rule [420].

1.2.3 Associative Testing Functions

Complementarily to training functions, multiple testing schema have been also proposed for associative classifiers [122, 99, 130]. Given a training function and a new observation, its labeling is typically accomplished by recovering the regions of the learned rules whose pattern φ_B best matches the values of the new observation and computing the strength of each class. The strongest condition is outputted as the estimated class if we want a deterministic output, otherwise the relative strength per class can be seen as its probabilistic value. Illustrating, CMAR [401] retrieves subspaces with exact matching and computes the classes' strength using a weighted χ^2 (WCS) calculus.

However, in many settings the exact matching criterion is restrictive as it can lead to a small (possibly empty) subset of regions and neglects the contributions of regions with good but non-exact matchings. To tackle these problems, previous work by Henriques and Antunes [321, 302] proposes relaxations on the matching functions (allowing, for instance, the presence of shifts and the matching of a subset of overall values), as well as weights to penalize non-exact matchings [302]. In these contexts, the rule score is proportionally affected by the extent of the shift or by the percentage of matching values (above a minimum threshold). Lazy classifiers that retrieve classification rules once a new (testing) observation is given have been also proposed to address this challenge [649, 122].

1.3 Solution

Learning of effective associative classifiers is driven by three major requirements:

- effective *discovery of biclusters*, implying the search for a flexible structure of (dissimilar) biclusters, each satisfying specific homogeneity, significance and discriminative criteria;
- effective *composition of biclusters* (training function);
- effective *matching of biclusters* (testing function).

To satisfy these requirements, we propose BiC (Biclustering-based Classifier). Accordingly, BiC: 1) extends the proposed biclustering algorithms to guarantee an effective and efficient discovery of discriminative biclusters only (*Section 1.3.1*); 2) relies on state-of-the-art training functions with revised scoring schema to adequately test biclusters with varying tolerance to noise and coherency strength (*Section 1.3.2*); and 3) defines testing functions that minimize the problems associated with the scarcity of matching and jeopardizing of scores (*Section 1.3.3*).

1.3.1 Discovery of Discriminative Biclusters

In order to guarantee the relevance of the discovered regions, BiC extends the proposed pattern-based biclustering algorithms proposed throughout *Book III* towards labeled data contexts. Three reasons justify this choice:

- pattern-based biclustering provides optimality guarantees, allowing the exact quantification of their the impact for learning classification models from high-dimensional data contexts.
- pattern-based biclustering offers flexible and parameterizable biclustering models, where the coherency (both coherency strength and assumption), quality (tolerance to noise) and statistical significance can be easily affected. As such, this enables the optimization of the properties of the underlying regions to improve the behavior of the target associative classifiers;
- pattern-based biclustering is not prone to discretization problems, handle arbitrary levels of sparsity, can effectively incorporate (domain-driven) constraints to guide the search and, among others, accommodate principles to guarantee the scalability of the searches.

To guarantee an optimal trade-off among flexibility, optimality and efficiency, these algorithms establish a formal link between biclustering and pattern mining (including frequent itemset mining, association rule mining, sequential pattern mining and graph mining). In this context, the pattern-based notions of support (number of observations) and pattern length (number of features) is applicable to the target regions, as well as confidence and interestingness metrics applied over decision rules. A consistent integration of these algorithms, seizing efficiency gains from their combined used, was described in *Chapter III-9*.

BiC extends these integrative pattern-based biclustering searches for each class-conditional data partition (Figure 1.3). In this context, it returns $|C|$ sets of biclusters, where $|C|$ is the number of classes. The default parameterizations of these algorithms (largely discussed in [310, 311, 303] and throughout *Book III*) were preserved in BiC.

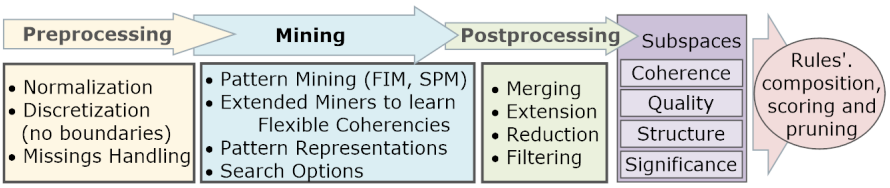


Figure 1.3: Class-conditional discovery of pattern-based biclustering models.

1.3.1.1 Multi-Label Discriminative Criteria

Under the proposed strategy, BiC is only able to deliver class-conditional biclusters with (possibly) significant support. In order to guarantee their discriminative power different strategies can be considered. First, a bicluster can be considered discriminative if its pattern $\varphi_{\mathbf{B}}$ (Def.III-1.1) is found to be frequent only in the context of a single class. In other words, a class-conditional bicluster is discriminative if similar biclusters are not found for the remaining classes. Understandably, this criteria cannot detect biclusters that are not able to discriminate an isolated class, yet are able to discriminate a group of classes.

Second, a bicluster can be considered discriminative if its pattern $\varphi_{\mathbf{B}}$ is found to be infrequent for at least one class. Although this view addresses the previous problem, it suffers from an additional drawback. Illustrating, consider a dataset with three classes (each labeled on 100 observations), and 0.4 to be the minimum fraction of observations from a class-conditional data partition that guarantee that a pattern is frequent. In this context, if a given pattern $\varphi_{\mathbf{B}}$ shows a support of $\{sup_{\mathbf{B}}|c_1{=}38, sup_{\mathbf{B}}|c_2{=}43, sup_{\mathbf{B}}|c_3{=}41\}$, according to the previous formulation would be considered discriminative for $\{c_2, c_3\}$, although $\varphi_{\mathbf{B}}$ is clearly non-discriminative.

Third, in order to address the previously depicted challenges, the discriminative power can be given by the confidence of a rule with multiple labels in the consequence, $conf_{\mathbf{B}\Rightarrow C}$ (where $C \subseteq C$) as defined in Def.1.3. In

this context, non-significant differences in support (even when associated with different frequency outcomes) are correctly interpreted. Given the previous illustrative case, since $conf_{\mathbf{B}\Rightarrow c_1} \approx conf_{\mathbf{B}\Rightarrow c_2} \approx conf_{\mathbf{B}\Rightarrow c_3}$, $\mathbf{B}$ is correctly interpreted as non-discriminative. However, the confidence (and peer discriminative scores) can rapidly increase for a combination of classes in rule's consequent. In this context, comparing the confidence of different rules is prone to errors related with the length of the consequent (e.g. $conf_{\mathbf{B}\Rightarrow\{c_2,c_3\}} \gg conf_{\mathbf{B}\Rightarrow c_2}$).

Def. 1.3 Given a labeled dataset $\mathbf{A}$, and a bicluster $\mathbf{B}=(\mathbf{I},\mathbf{J})$ with coherence across rows and pattern $\varphi_{\mathbf{B}}$ (expected values in the absence of adjustment and noise factors according to Def.1.1), then the pattern **support**, $sup_{\varphi_{\mathbf{B}}}=|\mathbf{I}|$, is the number of observations respecting φ_B. Given a set of labels C in C, the support of C, sup_C, is the number of observations with a label in C. Given a decision rule R:$\mathbf{B} \Rightarrow C$, the **rule support**, $sup_R = sup_{\varphi_{\mathbf{B}}|C} = \Sigma_{c_i \in C} sup_{\varphi_{\mathbf{B}}|c_i}$, is the number observations respecting φ_B with a label in C, and the *rule confidence*, $conf_R = conf_{\varphi_{\mathbf{B}}|C} = \frac{sup_{\varphi_{\mathbf{B}}|C}}{sup_{\varphi_{\mathbf{B}}}}$, is the fraction of C-conditional observations from $\varphi_{\mathbf{B}}$ supporting observations.

Although this problem can be minimized by comparing the confidence of rules with the similar consequent's length, it suffers from an additional problem. Confidence is not able to adequately deal with imbalanced data.

Def. 1.4 Given $\mathbf{A}$, a set of labels C in C and a decision rule $R : \varphi_{\mathbf{B}} \Rightarrow C$ in $\mathbf{A}$, the **lift** of a rule is $lift_R = \frac{sup_R}{sup_{\varphi_B} sup_C}$.

Finally, and with the goal of addressing previous problems, the lift of a rule (Def.1.4) can be used as the discriminative criterion since it normalizes confidence by the support of the labels in the consequent. Lift is preferable over commonly used discriminative power scores since it deals with the: 1) structural data imbalance associated with the number of observations per class, and 2) induced imbalance associated with the creation of rules with disjunctive consequents. Revisiting the illustrated scenario, given R_1:$\mathbf{B} \Rightarrow \{c_2\}$ and R_2:$\mathbf{B} \Rightarrow \{c_2, c_3\}$, then $lift_{R_1} = conf_{R_1}/sup_{R_1} \approx \frac{1}{3}/\frac{1}{3} \approx 1$, $lift_{R_2} = conf_{R_2}/sup_{R_2} \approx \frac{2}{3}/\frac{2}{3} \approx 1$, clearly indicating that these rules have no discriminative power. The higher the lift, the higher the discriminative power. A lift close or inferior to 1 indicates a loose discrimination.

1.3.1.2 Dealing with Noisy Regions

Understandably the previously introduced notions are only valid if the inputted biclusters are perfect, i.e. do not tolerate noise. Although the introduced support and confidence have been largely applied in the context of associative classification (due to its original orientation towards transactional data), these notions cannot consider the presence of noisy elements and of slight deviations on the pattern expectations. Illustrating, assuming the presence of a bicluster pattern $|\varphi_{\mathbf{B}}|=10$ and class-conditional data partition where all the rows match 9 from the 10 elements from $\varphi_{\mathbf{B}}$ (i.e. each row has one arbitrary noisy element). In this context, the class-conditional support of this pattern is incorrectly interpreted as 0.

To address this problem, we propose a variant of the introduced support notion in order to count the observations with noise (percentage of values not satisfying a given coherence strength) below a specific threshold. As this noise-sensitive counting affects the support calculus, it consequently affects the lift calculus. This guarantees an adequate discriminative view. Defs.1.5 and 1.6 below define these notions.

In this context, in order to compute these weighted metrics, linear matchings can be performed for each bicluster and the partitions where the bicluster is absent. Empirical evidence shows that when guaranteeing the dissimilarity between the outputted set of biclusters, the computational overhead associated with this step is not significant. Note that this contrasts with the principles proposed in the context of *Section 5*, where the absence of dissimilarity guarantees led to the definition of more scalable searches.

Def. 1.5 Given a coherence strength δ and noise threshold ϵ, an observation x_i respects φ_B if:

$$\frac{\kappa_i}{|J|} < \epsilon, \quad where\ \kappa_i = \Sigma_{j \in J} \eta_{ij} \ \ and\ \eta_{ij} \in [-\delta/2, \delta/2]) \tag{1.1}$$

Given a pattern $\varphi_{\mathbf{B}}$, its **weighted support** is $sup_{\varphi_{\mathbf{B}}} = \sum_{i \in I \wedge \kappa_i \| J| < \epsilon} (|J| - \kappa_i)^a$, where a=2 by default.

The proposed weighted support can be parameterized with a factor that determines how the level of noise may impact the computed support. In this context, linear, squared (default) or other penalizations can be easily specified.

Def.1.6 extends the traditional view of lift from transactional data towards real-valued biclusters possibly following non-constant coherencies and arbitrary levels of noise. In this context, lift reveals the noise-sensitive strength of the rule weighted by the representativity of the labels in the consequent.

Def. 1.6 Given a set of labels C in $\mathcal{C}$ and a decision rule $R : \varphi_{\mathbf{B}} \Rightarrow C$ in A, let the *support* of a rule, $sup_R = sup_{\varphi_B|C}$, be the number of (possibly noisy) observations respecting φ_B with a label in C. Accordingly, the weighted confidence and **weighted lift** of a rule are $conf_R = sup_R / sup_{\varphi_{\mathbf{B}}}$ and $lift_R = sup_R / (sup_{\varphi_B} sup_C)$, where both sup_R and $sup_{\varphi_{\mathbf{B}}}$ are computed according to Def.1.5.

A final problem is related with the fact that when a bicluster does not satisfy a minimum frequency criteria for a given partition, it is not modeled, and thus its support is unknown. The absence of support calculus for certain class-conditional partitions is undesirable as it prevents an efficient scoring rules. To alleviate this problem, different strategies can be applied, including constraint-guided searches in order to efficiently discover a bicluster with apriori known pattern $\varphi_{\mathbf{B}}$ on certain partitions through the use of succinct constraints. This strategy can follow similar principles to the strategies proposed in *Section 7.3*.

1.3.1.3 Efficient Learning of Rules with Disjunctive Consequents

As we previously described, we allow for the composition of decision rules with disjunctions of labels in the consequent (according to Def.1.1) to be able to include relevant regions that discriminate groups of labels. These rules are efficiently generated based on two principles.

First, $\varphi_B \Rightarrow c_1$ and $\varphi_B \Rightarrow c_2$ rules are only candidates for joining consequents when each have promising (yet non-significant) levels of discriminative power. In this context, the resulting rule $\varphi_B \Rightarrow \{c_1, c_2\}$ have to necessarily show an improvement on the computed lift against both of previous rules. Rules with n-wise consequents are then compared with rules with a single label to compose rules with $(n+1)$-wise consequents.

Second, the components associated with the lift calculus are maintained for each rule. This guarantees that the computation of the lift for the new rule is directly derived from the (intermediate) scores of the candidate rules.

1.3.2 Training

BiC makes available state-of-the-art structures of association models that provide an efficient way to navigate across rules. The simplest structure is an ordered set of tuples (pattern, coherency, classes, score).

By default, and similarly to CMAR [401], BiC composes a tree structure where rules are inserted according to their priority based on the support, length and discriminative power.

Additionally, BiC penalizes biclusters that are similar to other biclusters with higher priority. Similarity is computed using the generalized Jaccard-index on the shared features. Note, however, that the application of these penalizations have a slight impact since the discovered class-conditional biclusters already have guarantees of dissimilarity (according to postprocessing principles proposed in *Chapter III-7*).

1.3.2.1 Integrative Score

An effective scoring of rules is essential to rank and prune rules in order to guarantee a balanced number of distinct rules per class or group of classes. We propose an integrative score combining the rule's discriminative power (using the proposed noise-sensitive lift), pattern length and pattern support.

Def. 1.7 Given a labeled dataset **A** and a rule $\varphi_{\mathbf{B}} \Rightarrow C$ in **A**, the proposed **integrative score** is defined as:

$$\omega_R = \alpha_1 \frac{sup_R}{sup_{\varphi_B}} \frac{sup_C}{sup_C} + \alpha_2 \frac{sup_R}{n} \frac{sup_C}{sup_C} + \alpha_3 \frac{|\varphi_B|}{m}, \tag{1.2}$$

where the first component corresponds to the discriminative power of the rule (given by its lift), the second component to its relative pattern support, and the third component to its relative pattern length. Accordingly, $\{\alpha_1{=}0.6, \alpha_2{=}0.2, \alpha_3{=}0.2\}$, from empirical evidence[a].

[a] Conducted sensitivity analysis described in Section VI-2.2.4

The integration of these interestingness criteria is critical to effectively rank regions according to their relevance. Additionally, it overcomes the typical problem related with the prioritization of small (often non-significant) biclusters by existing associative classifiers, resulting from an overemphasized focus on the confidence of the rules.

1.3.3 Testing

In the testing stage, the learned discriminative associative model is used to label new observations by identifying the rules that better match the observed values. From the set of matched rules, BiC adapts state-of-the-art calculus to compute the strength of each class [401, 302]. In particular, below we show how the weighted-χ^2 can be extended to effectively deal with disjunctions of labels on rules' consequent:

$$weighted\chi^2(c) = \sum_{match(\varphi_B \Rightarrow C | c \in C)} \frac{sup_c}{sup_C} \frac{(\chi^2_{\varphi_B})^2}{MCS}, \quad with \;\; MCS = (min\{sup_{\varphi_B}, sup_C\} - \frac{sup_{\varphi_B} sup_C}{N})^2 N \times e, \tag{1.3}$$

where N is the number of matches and e is defined as $\frac{1}{sup_{\varphi_B} sup_C} + \frac{1}{sup_{\varphi_B} N - sup_C} + \frac{1}{N - sup_{\varphi_B} sup_C} + \frac{1}{N - sup_{\varphi_B}(N - sup_C)}$.

1.3.3.1 Computing Class Strength

In order to benefit from the proposed integrative score, we propose a new calculus to determine the strength of each class, referred as weighted integrated score (*WIS*). Given a testing observation, this calculus is based on three simple steps. First, the set of matched rules with the testing observation are identified. Second, their priority is computed according to Def.1.7. Third, whenever the matched rules are associated with more than one label C, a proportional adjustment is applied to correctly weight the influence of each individual class $c \in C$ based on their relative support. Finally, the weighted scored from the matched rules per class are summed and outputted.

Accordingly, the *WIS* for a specific class $c \in \Sigma$ is by default given by:

$$WIS_c = \sum_{match(R:\varphi_B \Rightarrow C | c \in C)} \frac{sup_c}{sup_C} \omega_R \tag{1.4}$$

The strongest class, $c \in C$, is outputted as the estimated class, if we want a *deterministic output*. Otherwise, the strength of each class is computed, normalized and returned as the *probabilistic output*.

1.3.3.2 Matches and Applicable Relaxations

Matching occurs if the values of the testing observation respect a bicluster pattern (Def.1.5). Yet, even in the presence of a large number of biclusters, the probability of matches to occur can be considerably low.

Thus, the introduction of relaxations is critical to consider matchings when a testing observation respects the majority (but not all) of the expected values from a bicluster pattern. In this context, we provide a simplistic

calculus to guarantee that the overall matching error is below a given threshold.

Def. 1.8 Given a rule $R : \varphi_B \Rightarrow C$ with score ω_R and a matching threshold θ, an observation $\mathbf{x}_{new}$ ***matches*** φ_B if it respects φ_B ($\frac{\kappa_{new}}{|J|} < \theta$, where κ_{new} is given by Def.1.5) with *score* $\omega_R \times (|J|-\kappa_i)^a$, where a=2 by default.

From empirical evidence, BiC uses, by default, 80% as the minimum percentage of matched values and uses a quadratic penalization of the score based on the percentage of mismatches. Alternative penalizations can be easily considering by parameterizing a in Def.1.8. These criteria guarantees a sound similarity for matches to occur, yet allows a parameterizable degree of mismatches to guarantee the presence of a medium-to-large matches per testing observation in order to guarantee that classification decisions can be made with higher confidence.

1.4 Results and Discussion

Results are organized as follows. First, we compare the performance of BiC against state-of-the-art classifiers on synthetic data with varying properties. Second, we extend this analysis towards high-dimensional biological and social data. The proposed classifiers[3] were implemented in Java (JVM v1.6.0-24). In particular, we compare BiC against associative classifiers based on pattern mining (using CMAR [401] after adequate data itemization according to the input δ) and biclustering (using FDCluster [661]), as well as against 3 non-associative classifiers prepared to learn from high-dimensional data without the need to rely on feature selection: support vector machines (SVM), Bayesian networks (BayesNet) and multivariate discriminants from Weka [286]. The experiments were run in an Intel Core i5 2.80GHz with 6GB of RAM.

Further empirical evidence for the relevance of the proposed contributions is provided with an increased detail in the experimental sections of the upcoming chapters (in which BiC's behavior is extended and revised).

1.4.1 Results on Synthetic Data

To assess BiC we make use of the synthetic data generator proposed in *Section II-3.3*, which is able to simulate regularities commonly observed in labeled biological data. In particular, the generated data combine global and local regularities according to the following parameters:

- data size and dimensionality, number of classes and class imbalance degree;
- mixture of class-conditional multivariate distributions (global regularities);
- number, coherency, noise and plaid effects of planted biclusters (local regularities);
- Uniform or Gaussian distributions to define the discriminative power, support and length of biclusters;

Table II-3.4 describes the parameters used to generate the synthetic datasets. Illustrating, biclusters with varying discriminative power were planted by generating patterns with a distinct number of supporting observations for different classes based on the expected confidence $\mu(\phi) \in \{90\%,80\%,70\%,60\%\}$ (with $\sigma(\phi)$=2%) per bicluster. Results are an average of 20 instances per setting.

Figure 1.4 validates whether BiC is able to accurately perform classification in synthetic data with varying properties (default settings according to Table II-3.4). In particular, we measure its ability against peer classifiers to model discriminative regions with varying: 1) coherence strength (from δ=5% to δ=50%), 2) number of supporting observations (from Uniform distributions with the expected value varied from 20% to 50%), 3) amount of noise (from 0% to 10%), and 4) discriminative power (from 60% to 100%). Results show Bic distinct superiority against associative classifiers (CMAR [401] and FDCluster [661]) and global classifiers from Weka [286] (support vector machines, Bayesian networks and multivariate discriminants). In particular, the provided analysis show the importance of the proposed discovery, training and testing functions to learn from regions with varying size and coherency strength, to robustly handle noise, and to model regions with looser levels of discriminative power in the absence of highly discriminative regions.

[3] Available in http://web.ist.utl.pt/rmch/software/bclassifier

Three additional observations are retrieved. First, non-associative classifiers are not inherently prepared to identify discriminative patterns supported by a small subset of overall observations. As such, their performance levels are only acceptable when at least 50% of observations from a given class support a discriminative pattern. Furthermore, their performance rapidly deteriorates when the planted profile is not distinctively discriminative (confidence below 75%). Second, traditional associative classifiers based on patterns derived from frequent itemsets show the worse performance. In particular, the observed differences in the performance against the remaining approaches are mainly driven by their inability to accommodate noise (with impact on both the discovery and testing stages). Third, although classifiers based on discriminative biclusters show improvements against pattern-based classifiers, their performance is still inferior to BiC. We hypothesize that three major reasons justify these differences: 1) adequacy of the proposed integrative scores, 2) ability to include disjunctions of labels in rules' consequent, and 3) relaxations to guarantee an adequate number of testing matches.

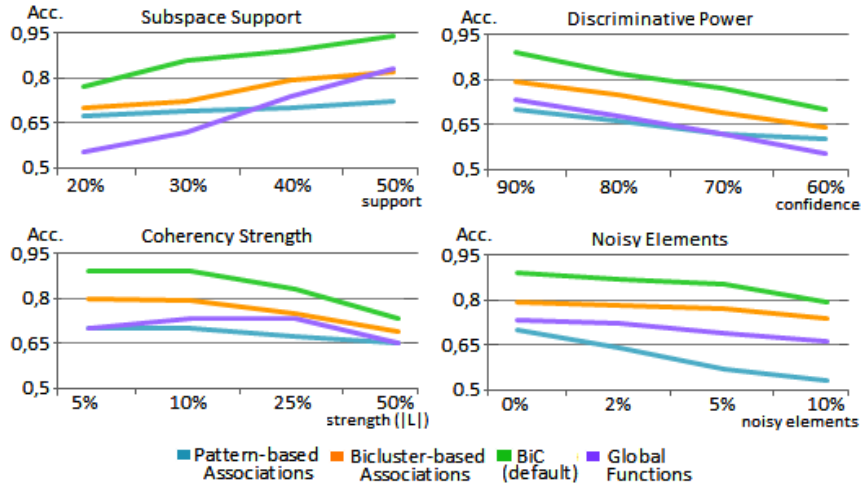


Figure 1.4: BiC's ability to learn from discriminative biclusters against peer classifiers in the presence of regions with varying support, discriminative power, coherency strength, and noise.

Figure 1.5 further enlarges the previous analysis in order to evaluate the accuracy and efficiency of BiC for data with varying size and dimensionality. First, the accuracy analysis shows the ability of BiC to consistently preserve high accuracy levels across data settings. Second, although SVMs and discriminant functions are not well prepared to deal with the generated datasets, Bayesian networks show an interesting and contrasting property: as the area of regions grow for larger data settings, their discriminative effects become more easily identified, leading to an improved accuracy. Third, the efficiency analysis show that BiC is a competitive option with peers (assuming the absence of scalability principles and its parameterization with flexible coherency assumptions), being able to learn from a (n=2000,m=5000)-space in less than a minute.

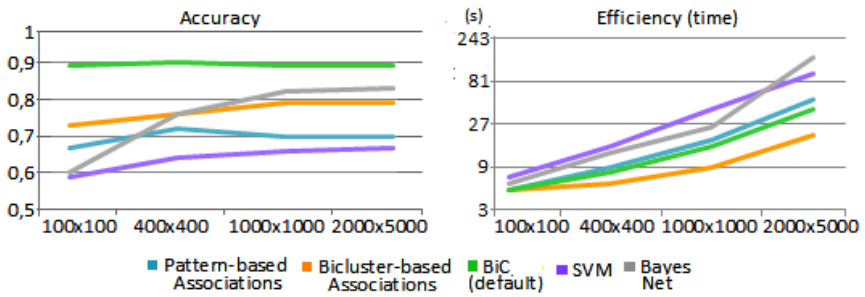


Figure 1.5: Accuracy and efficiency of Bic against state-of-the-art classifiers over synthetic data.

1.4.2 Results on Real Data

We selected 8 real datasets: 4 from biological contexts and an additional set of 4 from social contexts. The biological datasets are (high-dimensional) labeled expression data[4] for the classification of: *1)* distinct types of lymphoma (m=4026 *features*); *2)* leukemia (m=7129); *3)* embryonal tumours outcome (m=7219); and *4)* colon cancer (m=2000). Two of the four datasets from social contexts correspond to the collaborative filtering of items from the Jester recommender system[5], respectively m=100 items (jester D_1) and m=150 items (jester D_2) rated using a continuous [-10,10] scale. Only users rating over 80% of the overall items were considered (n=17.438 for D_1 and n=9.542 users for D_2). Non-rated items are interpreted as missing values by the applied classifiers. We selected the last five items rated by all users from each dataset as 5 sets of trinary classes for prediction ($\{0|a_{ij}\in[-10,-2[, 1|a_{ij}\in[-2,2], 2|a_{ij}\in]2,10]\}$) and report the average performance of classifiers across such items. Finally, we used two datasets from psychological questionnaires with grading questions (integers from 1 to 5) [6]: Cattell's test (referred as 16PF) with m=163 grades answered by n=49159 individuals and Simon Baron-Cohen's test (referred as EQSQ) with m=120 grades answered by 13256 individuals. For these datasets we aim to predict the self-rated accuracy ({medium,high,very-high}) based on the assessed profile.

An in-depth description of discriminative regions learned by BiC for each one of these datasets is provided with detail in the next chapter. Below, we center our analysis on the gathered accuracy and sensitivity levels. For this analysis, BiC was parameterized with BicPAMS in order to be able to discover non-constant coherencies (according to *Chapter III-9*). The gathered results (and their variance) were collected from 10 cross-fold validation. The sensitivity was computed from the case class for the biological datasets and from the less-accurate class for the social datasets.

Figure 1.6 shows the accuracy and sensitivity levels of BiC against associative and non-associative classifiers. Results confirm the superiority and critical role of using BiC's searches with flexible biclustering models. In fact, most of the observed differences are statistically significant (t-Student tests with 9 degrees of freedom and p-value<0.05). We hypothesize that this is not only a result of the tackled limitations of existing associative classifiers (adequate discriminative, scoring and matching criteria) but mainly driven by BiC's ability to model regions with varying coherency strength, coherency assumptions and quality. This hypothesis is further motivated by the fact that, unlikely BiC, the analysis of peer associative classifiers reveals that they are only able to find a few discriminative regions.

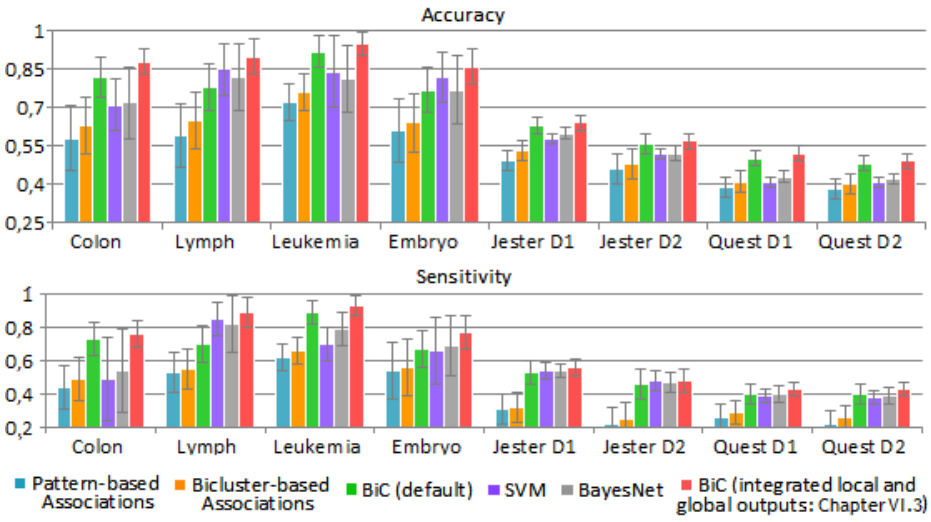


Figure 1.6: Comparison of the performance of BiC against associative and global classifiers over real data.

[4] http://eps.upo.es/bigs/datasets.html
[5] http://eigentaste.berkeley.edu/dataset/
[6] http://personality-testing.info/

1.5 Conclusion and Implications

In this chapter, we motivated the problem of learning classification models from labeled high-dimensional data, often characterized by the presence of local regularities associated with discriminative regions. Associative classification was introduced as a means to surpass the biases from the application of distinct forms of dimensionality reduction. In this context, the criticisms of existing associative classifiers preventing its broader application were carefully surveyed. To address these criticisms, the contributions and limitations of available research were surveyed, and the requirements for the learning of effective associative classifiers enumerated according to the: *1)* discovery of discriminative and informative regions biclusters with parameterizable properties and dissimilarity guarantees; *2)* adequate scoring and compositions of these regions; and *3)* robust testing of new observations against the composed model. A new associative classifier, BiC, was proposed to satisfy these requirements.

BiC provides three major groups of contributions. First, BiC relies on extended information-gain scores to adequately identify discriminative rules from imbalanced data and with disjunctions of labels in the rules' consequent. Second, BiC is able to efficiently identify and adequately score biclusters able to discriminate a group of classes (but not a single class). Third, BiC considers noise-tolerant support criteria to guarantee the adequate scoring of rules. Furthermore, it adequately combines discriminative power, support and length within a single integrative score in order to have that some rules unfairly jeopardize the learning. Fourth, BiC provides effective testing functions that guarantee: 1) an adequate number of matches per testing observation based on similarity relaxations and 2) a new robust calculus of classes' strength.

Results on both real and synthetic high-dimensional datasets reveal that the performance of BiC is associated with statistically significant gains, both against peer associative classifiers and global classifiers. The collected empirical evidence positions BiC as a promising approach to learn from high-dimensional data.

Future Work. In this context, three major directions for future are identified. First, an in-depth analysis of the properties of discriminative regions from labeled biomedical and social data domains. Second, the parameterization BiC to systematically study the impact that coherency assumptions, noise tolerance and statistical significance has on the performance and properties of the learned associative models. Third, we expect to use this knowledge as an input to dynamically adapt the behavior of state-of-the-art classifiers.

2

Classification from Regions with Non-Constant Coherency

Although existing associative classification models generally rely on constant regions given by discriminative patterns, many meaningful regions in biomedical and social data are given by non-constant coherencies, including additive, multiplicative, symmetric, order-preserving and plaid coherencies. The discovery of these regions is critical to identify, for instance, regulatory modules or groups of individuals with non-trivial (yet coherent) behavior [257, 310, 334].

Furthermore, the discovery of regions with flexible coherence can be used to minimize the existing criticisms of associative classifiers, namely: 1) guarantee a more diverse set of regions that guarantees a more adequate space exploration and minimizes the scarcity of matchings; and 2) promote the discovery of larger regions (an implication of learning from more regions with more flexible coherency assumptions). As a result, the learning becomes less susceptible to underfit the input data.

Despite the importance of discovering and learning from discriminative regions with non-constant coherencies in noisy and real-valued settings, there are not yet attempts towards this end [660, 497, 99]. This observation leads us to the two target research questions by this chapter: How to effectively shape (associative) classifiers to learn from regions with flexible coherency? How does their performance varies with the properties of the modeled biclusters?

Learning adequate classifiers from non-constant regions has challenges across their three major steps: 1) the discovery of (dissimilar) discriminative regions with flexible coherencies is non-trivial, 2) training is challenged by the fact that more flexible coherencies are associated with larger regions that can unfairly score higher and hamper the learning, and 3) existing testing functions are not able to match non-constant biclusters.

As such, this chapter addresses the introduced challenges by proposing new principles for learning associative classifiers from discriminative biclusters with flexible coherence assumptions and parameterizable tolerance to noise. Follows the five major contributions of this chapter:

- extended discovery of discriminative regions to efficiently model (dissimilar) biclusters with varying quality and additive, multiplicative, order-preserving and plaid coherency assumptions;
- a new penalization schema for non-constant coherencies based on their degree of flexibility (for preventing that biclusters with flexible coherencies jeopardize the learning);
- extended integrative score to better weight the quality of a region (deviations from pattern expectations);
- extended testing score to compute the strength of each class from multiple interestingness criteria (lift, support, quality, length and coherency);
- new matching criteria to test observations against non-constant biclusters (similarity assessment sensitive to the allowed adjustment factors).

As a result, BiC is extended into a new associative classifier, FleBiC (Flexible Biclustering-based Classifier), to learn an effective composition of biclusters with multiple coherencies and classify observations using noise-tolerant and coherency-sensitive scores.

Results on synthetic and real data demonstrate the relevance of modeling non-constant biclusters to enhance the

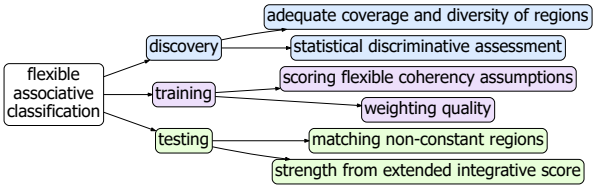


Figure 2.1: Contribution for learning associative classifiers from regions with flexible homogeneity.

performance of classifiers and show FleBiC's ability to unravel non-trivial and meaningful discriminative relations from real-data contexts. In particular, we quantify and statistically assess the gains of FleBiC against BiC and other classifiers, and analyze non-constant regions retrieved from biomedical and social data domains.

This chapter is organized as follows. *Section 2.1* provides the background, motivating the need and current limitations for learning flexible associative classifiers. *Section 2.2* describes FleBiC. *Section 2.3* gathers observations from assessing FleBiC and analyzing the learned associative models. Finally, the major contributions and implications are synthesized.

2.1 Background

A bicluster can have coherence of values across observations, features or overall elements, where the values typically follow a constant model [429]. When observations are labeled, a coherency orientation across observations should be preferred for the learning of decision models. Although discriminative biclusters with coherency across features can be meaningfully discovered, their use for testing new observations is less trivial. Complementarily to coherency orientation, additional coherency assumptions can be given by additive, multiplicative and symmetric, plaid and order-preserving [59, 303].

According to Def.II-3.1, let the elements in a bicluster $a_{ij} \in (\mathbf{I},\mathbf{J})$ have *coherence across observations* given by $a_{ij}{=}k_j{+}\gamma_i{+}\eta_{ij}$, where k_j is the expected value observed for feature j, γ_i is the adjustment for observation i, and η_{ij} is the noise factor. Given a specific coherence strength $\delta \in [0, max_A{-}min_A]$, $a_{ij}{=}k_j{+}\gamma_i{+}\eta_{ij}$ where $\eta_{ij}{\in}[{-}\delta/2,{+}\delta/2]$.

According to Def.II-3.2, the γ factors define the coherence assumption: *constant* when $\gamma{=}0$; *multiplicative* if a_{ij} is better described by $k_i\gamma_j + \eta_{ij}$; and *additive* otherwise. An *Order-preserving* assumption is verified when the values of features induce the same linear ordering across observations. *Symmetries* can be considered by re-describing the data elements by $b_i \times a_{ij}$ where $b_i \in$ {-1,1}. A *plaid* assumption considers the cumulative effect of the contributions from multiple biclusters on areas where their rows (observations) and columns (features) overlap.

For these coherency assumptions, the *bicluster pattern* $\varphi_{(I,J)}$ is given by the set of expected values in the absence of adjustments and noise $\cup_{j=0}^{|J|}\{k_j\}$. Consider the illustrative additive bicluster $(\mathbf{I},\mathbf{J}){=}(\{\mathbf{x}_1,\mathbf{x}_2\},\{\mathbf{y}_1,\mathbf{y}_2,\mathbf{y}_3\})$ in $\mathbb{N}_0^+$ with coherence across observations, where $(\mathbf{x}_1,\mathbf{J}){=}\{1,-3,2\}$ and $(\mathbf{x}_2,\mathbf{J}){=}\{-3,5,-4\}$. Assuming $\delta{=}1$, this bicluster can be described by $a_{ij}{=}b_ik_j + \gamma_i$ with the pattern $\varphi{=}\{k_1{=}0,k_2{=}$-2,$k_3{=}1\}$, supported by two observations with additive factors $\gamma_1{=}1$ and $\gamma_2{=}3$ and a symmetry on the second observation $b_2{=}{-}1$.

High-dimensional biomedical and social data is characterized by the presence of biclusters with flexible coherence [303, 429]. Table 2.1 motivates the relevance of selecting regions with flexible coherence by highlighting biomedical and social contexts where their discovery is critical for learning tasks.

Many biclustering algorithms emerged in the last decade to address this need (Table 2.2). Yet, since they were originally proposed for learning descriptive models, they have inherent limitations that prevent their adequate use for (associative) classification. In particular, these attempts are neither able to discover flexible biclustering models (restrictions on the placed structures, coherency strength, allowed coherency assumptions, tolerated noise and statistical significance) nor able to provide optimality guarantees, preventing an adequate assessment of how the properties of the underlying regions affect the behavior of the target associative classifiers.

Nevertheless, and despite the availability of biclustering algorithms able to model non-constant coherencies

Coherence	*Illustrative biclusters across biomedical and social domains*
Additive and Multiplicative	Coherencies used to allow the occurrence of shifting and scaling factors across observations (Figure III-1.2). Illustrating, two genes may be regulated in the same subset of conditions (features) but show different expression levels explained by a shifting or scaling factor associated with their distinct responsiveness, or the bias introduced by the applied measurement and preprocessing [310]. These factors are also critical to analyze physiological and clinical data to handle the structural differences across individuals [122]. In social domains, these factors are relevant to model social interactions with coherent behavior but differing in the extent of frequency and popularity of actions, and to group subjects with identical variation of preferences during browsing and collaborative filtering [257].
Order-Preserving	Order-preserving biclusters were originally proposed to find genes co-expressed within a temporal progression (such as stages of a disease or drug response) [59]. Yet, they have been also largely applied in static biological contexts where gene expression or molecular concentrations coherently vary across samples [311]. This coherence can be also applied to: find sets of nodes in (social and biological) networks with an order-preserving degree of influence across another set of nodes; to support task planning and scheduling; and to discover order-preserving preferences from collaborative filtering data [311, 415]. Order-preserving biclusters can emulate constant, additive and multiplicative coherencies, leading to more inclusive solutions with larger and less noise-susceptible regions.
Symmetric	In biological contexts, symmetries are key to simultaneously capture activation and repression mechanisms within biological processes associated with biclusters in transcriptomic, proteomic or metabolic data [429]. In social contexts, symmetries are used to capture opposed (yet correlated) regularities associated with trading, tweeting, browsing and (e-)commerce activity [311]. Symmetries can be combined with the previous coherencies.
Plaid	Plaid models are essential to describe overlapping regulatory influence in biological contexts and cumulative effects in the interactions between nodes in social networks [393, 303]. Illustrating, consider a gene activated by a set of biological processes, a plaid coherence can consider their cumulative effect on the expression of a gene when more than one of these processes is active at a particular time. The plaid model can be also applied to study regulatory cascades, user behavior and trading operations, as these data contexts are also characterized by non-trivial influences between biclusters [303].

Table 2.1: Relevance of non-constant biclusters when learning from biomedical and social data contexts.

Coherence	*State-of-the-art algorithms*	*Limitations*
Additive and Multiplicative	Major attempts rely on merit functions based on variance, either more suitable to model additive factors (including residue-based approaches [134, 690]), or multiplicative factors (such as Fabia [324]). Some approaches unify these seemingly incompatible factors using linear geometry in hyper-spaces [235], evolutionary computing [532], and swarm intelligence [161].	1) Higher propensity to discover noisy constant biclusters instead of biclusters with strict additive and multiplicative coherence [310]; 2) Restrictions on the structure and quality of solutions.
Order Preserving	Greedy approaches iteratively discover-and-mask biclusters, including the pioneer OPSM [59] and its extension to model uncertain data with continuous distributions [209]. The few available exhaustive approaches, such as *u*Clustering [415], identify the largest regions respecting the ordering constraints, overcoming the quality and flexibility issues of the greedy peers.	1) Greedy solutions with restrictions on the structure (no overlaps) and no optimality guarantees; 2) Exhaustive approaches have efficiency bottlenecks and are highly susceptible to noise (perfect orderings only).
Symmetric	Some algorithms [616, 427] are able to combine constant coherencies with symmetries (also referred as sign-changes).	1) Not integrated with non-constant models; 2) Restricted to time series analysis (contiguous features).
Plaid	First algorithmic attempts rely on greedy searches that discover one bicluster at a time and subtract the respective contributions from data [393]. Generative alternatives to minimize some problems and learn the whole biclusters at a time were propose using expectation-maximization [573], binary matrix factorization based on non-parametric Bayesian models to better approximate the complex joint distributions of a plaid models [449, 113].	1) Exact additive model of contributions (their composition may not increase linearly in real contexts); 2) Require all the elements in the dataset to fit the plaid model; 3) Restrictions on the allowed type (generally constant) and number of biclusters.

Table 2.2: State-of-the-art contributions and limitations of algorithms to discover non-constant biclusters.

for nearly a decade, the existing associative classifiers based on discriminative biclusters are still focused on the discovery of biclusters with (approximately) constant values across observations [660, 497, 99].

The recent breakthroughs on biclustering proposed by the authors – BicPAM [310], BicSPAM [311], BiP [303] and BicNET [313] –, largely described through *Books III* and *V*, can however be used to tackle these limitations. Table III-9.1 from *Chapter III-9* delves point-by-point how the proposed contributions can be used to tackled the current limitations enumerated in Table 2.2. Furthermore, these algorithms can: *1)* accommodate meaningful constraints and relaxations, while seizing their efficiency gains; *2)* robustly handle different forms of noise (including the noise associated with the applied discretization); and *3)* effectively learn from sparse data.

As such, their use opens a new door to not only study the impact of using non-constant coherencies for associative classification, but going further on answering the target research question: How does the performance of (associative) classifiers varies with the properties (coherency, quality, significance, discriminative power) of the learned regions?

Def. 2.1 Given a (high-dimensional) tabular dataset $\mathbf{A}$ with labeled observations, let a decision rule be $R_k : \mathbf{B}_k \Rightarrow C_k$, where C_k is a subset of classes ($C_k \subset \Sigma$) and $\mathbf{B}_k$ is a region where observations have strong homogeneity (respecting *Def.II-3.1* and *Def.II-3.2* with parameterizable η_{ij}), significant support $P(\varphi_{B_k})$ and discriminative power $P(C_k|\mathbf{B}_k)$. The target task of **learning classifiers from non-constant regions** can be mapped as the task of: 1) discovering and composing decision rules $\{R_1, .., R_p\}$ derived from exhaustive structures of discriminative biclusters with flexible coherence (following additive, multiplicative, symmetric, order-preserving and plaid models) and parameterizable quality, and 2) matching observations $\mathbf{x}_{new} \in \mathcal{X}$ against these rules $M : \mathcal{X} \rightarrow C$.

2.2 Solution

To answer the target task (Def.2.1), we propose FleBiC (Flexible Biclustering-based Classifier), an extension of BiC. In this section, we organize the proposed contributions according to the major steps of FleBiC: *1)* discovery of flexible structures of discriminative biclusters with parameterizable coherence and quality (*Section 2.2*), *2)* effective composition of biclusters (*Section 2.2.1*), and *3)* effective scoring schema to test new observations (*Section 2.2.4*). Finally, *Section 2.2.1* analyzes the computational properties of FleBiC and discusses its parameterizations.

2.2.1 Discriminative Regions with Flexible Homogeneity

To enable the integrative discovery of flexible structures of biclusters with varying coherency assumptions (including constant, additive, multiplicative, symmetric, order-preserving and plaid models) and parameterizable quality, we rely on BicPAMS (see *Chapter III-9*). BicPAMS consistently combines the contributions of state-of-the-art pattern-based biclustering algorithms (including BicPAM [310], BicSPAM [311], BiP [303] and BicNET [313]) to enable the integrative search for biclusters with shifts, scales, symmetries, preserved orderings and plaid effects. Besides making use of the efficiency principles associated with each of the algorithms[1], BicPAMS is able to seize efficiency gains from the integrative search for regions with distinct coherency assumptions and from the integrated application of preprocessing and postprocessing steps (details in *Section III-9*). For the purpose of finding discriminative biclusters, BicPAMS is applied (as illustrated in Figure 1.3). FleBiC applies BicPAMS for each on each class-conditional data partition, returning $|C|$ biclustering models.

Dissimilarity. Understandably, the search for biclusters with varying coherency and tolerance to noise can result in voluminous outputs of biclusters, where subsets of biclusters may be highly similar. Understandably, this condition is undesirable as it can lead to incorrect decisions due to a biased scoring of testing observations related with matching with multiple similar biclusters. BicPAMS addresses this problem as it is able to adequately postprocess the discovered biclusters (with varying coherencies and quality) using efficient merging and filtering procedures. In this context, we also enhanced BicPAMS with principles to guarantee the diversity of biclusters. BicPAMS allows biclusters contained in larger biclusters as long as there are significant differences with regards to their sizes. This is an important principle to guarantee the output of biclusters with varying tolerance to noise and, mostly, to allow that biclusters with strict coherencies (generally included in an order-preserving bicluster) appear in the learned biclustering model. Similarly, BicPAMS allows biclusters with overlapping areas below a dynamically fixed threshold. Two biclusters may have a significant overlapping elements, yet their discriminative power be quite distinct. Since the knowledge regarding the discriminative power of a given bicluster may not be known a priori, dissimilarity guarantees are provided without too restrictive criteria. In this context, the computed scores in the training stage can be used to further prune the output set of biclusters and therefore augment their guarantees of dissimilarity.

Decision Rules. In order to learn classification rules from discriminative biclusters, FleBiC makes use of the principles implemented in BiC, including: 1) the weighted notion of support to adequately compute basic metrics for an observed pattern φ_B; 2) principles to guarantee the efficiency retrieval of the (noise-tolerantt) support; 3) weighted lift criteria to measure discriminative power; and 4) efficiency principles to compose decision rules with disjunctions of labels in their consequent. FleBiC further extends the discriminative criteria in order to guarantee the presence of a statistical test that can be used to filter non-discriminative rules. For this aim, FleBiC also performs χ^2 tests according to CMAR [401]. In this context, a multi-hypothesis correction is applied with a 0.1 confidence threshold. Understandably, in data contexts with a low number of decision rules that discriminate a particular class,

[1] Pattern-based biclustering enables exhaustive yet efficient pattern mining searches (including frequent itemset mining, association rule mining, sequential pattern mining, graph mining) by: 1) making use of anti-monotonic principles to narrow the search space and 2) introducing the possibility to parallelize algorithms, mine approximate patterns and/or rely on data partitioning strategies under certain optimality guarantees.

this threshold is either relaxed or the filtering bypassed. Independently of whether a particular region is rejected or not, the gathered p-value is used to affect the (integrative) score of the rule.

2.2.2 Training: Scoring Noisy and Non-Constant Regions

Extended Integrative Score of Rules. We extended the proposed integrative score to combine lift, pattern support and pattern length with two new criteria. First, the view on the rule's discriminative power given by the χ^2 tests. Second, a view on the quality of the underlying region in order to benefit matches with regions less susceptible to noise. In this context, we use the average deviation from the pattern expectations (as specified in *Section V-2.2*). With this criterion we guarantee a better balance between the size of regions and the tolerated noise, prioritizing larger regions as long as they have high quality.

Let Q be the quality of a bicluster either defined by the fraction of noisy elements (symbolic data), $Q=\frac{1}{|J|}\Sigma_{i\in I}\frac{\kappa_i}{|J|}$ (where κ_i is defined according Def.VI-1.5) or by the deviation against expectations (real-valued data). The integrated score is defined as:

$$\omega_R = \alpha_1(0.7\frac{sup_R}{sup_{\varphi_B}}\frac{sup_C}{sup_\Sigma} + 0.3\chi^2_{\varphi_B}) + \alpha_2(0.5\frac{sup_R}{n}\frac{sup_C}{sup_\Sigma} + 0.5\frac{|\varphi_B|}{m}) + \alpha_3 Q, \tag{2.1}$$

where $\{\alpha_1{=}0.6,\alpha_2{=}0.3,\alpha_3{=}0.1\}$, from empirical evidence (see *Section 2.2.4*).

Scoring Non-Constant Coherencies. A problem with the previous score is its inability to address the fact that different coherencies may show different degrees of flexibility. As illustrated in Figure 2.2, order-preserving biclusters have high flexibility degree as they are able to capture additive and multiplicative coherencies, which in turn are able to capture constant coherencies. Understandably, regions associated with coherency assumptions with higher flexibility have higher scores as they are typically associated with larger biclusters. In this context, biclusters with higher flexibility can jeopardize the learning since biclusters given by more restrictive coherencies become neglected. For this reason, it is important to introduce a new penalization weight, $\omega \times \nu$, for non-constant coherencies based on their degree of flexibility.

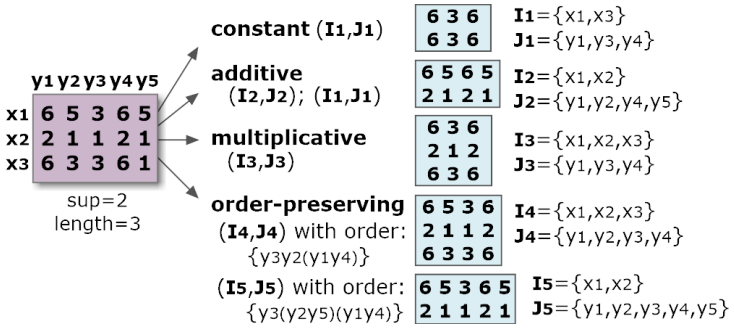


Figure 2.2: Varying degree of flexibility of non-constant biclusters.

From empirical evidence, the following penalizations are provided by FleBiC as default: order-preserving (ν=0.7 with symmetries and ν=0.75 otherwise), additive (ν=0.8 with symmetries and ν=0.85 otherwise), multiplicative (ν=0.9), and constant with symmetries (ν=0.95). When plaid effects are allowed, the penalization ν is given by the underlying coherency assumption in the absence of cumulative contributions from the overlapping biclusters.

2.2.3 Testing: Matching Observations against Non-Constant Regions

To determine if a testing observation respects a non-constant pattern we need to verify if the observed values can be described by an adjustment factor. Given a non-constant bicluster pattern $\varphi_\mathbf{B}$ and an observation $\mathbf{x}_{new}$, $\mathbf{x}_{new}$ matches $\varphi_\mathbf{B}$ if it can be described by $\varphi_\mathbf{B}$. In this context, three different settings are allowed: 1) $\gamma \neq 0$ (Def.II-3.1) can be

assumed to describe $\mathbf{x}_{new}$ in the presence of an additive, multiplicative or plaid pattern, 2) $\mathbf{x}_{new}$ values can differ from $\varphi_{\mathbf{B}}$ when $\mathbf{B}$ is described by an order-preserving assumption as long as the majority of ordering constraints are satisfied, and 3) b_i=-1 (Def.II-3.2) is allowed for $\mathbf{x}_{new}$ (symmetric values or reversed orderings) when $\mathbf{x}_{new}$ is tested against a symmetric pattern.

Illustrating, given a bicluster with pattern φ_B=$\{1.2,3.3,2.0\}$ on features $\{y_{89}, y_{459}, y_{892}\}$. If the bicluster is additive and a testing observation has values $\{3.1,5.3,4.1\}$ on the same features, the values are coherent under a shifting factor γ=2. If the same bicluster is order-preserving, the testing observation is also coherently described $(y_{89}<y_{892}<y_{459})$.

Complementarily to the matching criteria, the class strength calculus (VI-1.4) is also revised to accommodate the penalization factors associated with the observed coherencies, $WIS_c = \sum_{match(R:\varphi_B \Rightarrow C | c \in C)} \frac{sup_c}{sup_C} \nu \times \omega_R$

2.2.4 Algorithm

FleBiC (Algorithm 10) relies on the application of class-conditional biclustering using BicPAMS with decreasing support until a set with a minimum number[2] of 20 dissimilar, significant and discriminative biclusters is found per class whenever possible. Rules with disjunctions of labels in the consequent are generated from these discovered biclusters to compose the classifier according to the proposed scoring schema and weight penalizations to balance biclusters with distinct degree of flexibility. In the testing stage, the weighted integrated score is applied with relaxations to adequately match noisy regions with non-constant coherencies. In the presence of classification decisions with low-to-medium levels of confidence, the resulting score is integrated with the output of alternative (probabilistic) classifiers.

Computational Complexity. The computational complexity of FleBiC is bounded by the biclustering task, which depends on the size of the class-conditional matrix, distribution of values, and merging procedure (details in [310, 311, 303]). Scalability principles from pattern mining (data partitioning strategies and approximate searches), as well as the effective incorporation of constraints (based on user expectations and available background knowledge) can be used to guarantee the heightened efficiency of this step for high-dimensional datasets with a very high number of observations. The composition of rules, as well as training and testing steps are linear on the average number of features of the outputted sets of biclusters.

Parameterization of FleBiC. Although FleBiC is highly parameterizable, it can be effectively applied with either default or data-driven parameterizations. In this scenario, biclustering is applied with dynamically fixed behavior (according to [310, 311, 303]). The parameters associated with the training and testing functions are by default fixed according to a conducted a sensitivity analysis. For this end, we iteratively varied the combinatorial values of the controlled parameters (including but not limited to α_1, α_2, α_3) for synthetic data with varying properties (Table II-3.4) until the harmonic mean of the accuracy and sensitivity was maximized. Since the variations of the parameters' values were not significant across the different settings, we considered their average value as the default parameterization.

Nevertheless, for the purpose of understanding and improving the performance of the target associative classifiers based on the properties of the underlying regions, FleBiC's behavior can be easily parameterized. FleBiC provides the distinct possibility not only to parameterize the coherence criteria, quality and minimum support of biclusters, but also to control the minimum thresholds associated with their significance and discriminative power.

[2]Number fixed based on empirical evidence from a sensitivity analysis conduced over the datasets described in Table II-3.4.

Algorithm 10: FleBiC Core Steps

```
1  Training
   Input: data, /*remaining params dynamically fixed when absent*/ coherencies, PMiner
     stopCriteria /*min. disc. biclusters per class*/, discretizer, noiseHandler
2  begin
3      /* multi-symbol assignments to surpass discretization drawbacks [310] */
4      multiSymbolData ← discretize(data, discretizer, noiseHandler);
5      transDB ← createTransactions(multiSymbolData);
6      foreach c ∈ classes do
7          minSup ← 1;
8          {minFeatures, noiseAcc} ← findPatternExepctations(transDB[c]);
9          /* integrated BicPAM/BicSPAM/BiP searches */
10         do
11             biclusters[c] ← search(PMiner, c, transDB, minSup, coherencies);
12             /* significance tests and other ratios */
13             scores[c] ← computeWeightedScores(biclusters, transDB);
14             if stopCriteriaAchieved(stopCriteria, biclusters[c], scores[c]) then
15                 biclusters[c] ← merge(biclusters[c], noiseAcc);
16                 /* non-mandatory filtering and extension */
17                 biclusters[c] ← incDiscPower(biclusters[c], transDB, scores[c]);
18             minSup ← minSup×0.9;
19         while !stopCriteriaAchieved(stopCriteria, biclusters[c], scores[c]);
20     rules ← produceRulesWithDisjointLabels(biclusters, scores);
21     rules ← computeIntegratedScoreWeightedByCoherence(rules);
22     flebic ← composePriorTreeStructure(rules);
23     flebic ← CompactAndDissimilarRuleSets(rules);
24     return flebic;

25 Testing
   Input: observation, flebic, relaxation /*squared by def.*/, globalClassifiers /*optional*/
26 begin
27     /* matching depends on the coherencies of regions in flebic */
28     if maxNrClassMatches(observation, flebic)<2 then relaxation ← relax(relaxation); foreach c ∈ classes do
29         strength[c] ← computeWIS(observation, flebic, c, relaxation);
30     if maxVal(strength)<secondMaxVal(strength)×0.8 then
31         strength ← 0.4×strength+classDist(observation,globalClassifiers)×0.6;
32     return maxIndex(strength);
```

2.3 Results and Discussion

Results are organized as follows. First, we analyze the relevance of discovering discriminative non-constant biclusters and measure the impact of their use in the performance of associative classifiers. Second, we compare FleBiC with state-of-the-art classifiers. FleBiC[3] was implemented in Java (JVM v1.6.0-24) and tested over synthetic and real data using a 10-fold cross-validation on an Intel Core i5 2.80GHz with 6GB of RAM.

Data Settings. We preserved the synthetic datasets generated in the context of the previous chapter to evaluate the impact of the proposed extensions over BiC. We selected 8 real datasets: a) 4 biological datasets[4] for the classification of distinct types of lymphoma (m=4026 *features*), leukemia (m=7129), embryonal tumours outcome (m=7219), and colon cancer (m=2000), b) 2 collaborative filtering datasets from Jester recommender system[5] (with m=100 and m=150) to classify 5 attributes (we report their average) with three classes each; and c) 2 datasets from psychological questionnaires[6] with $|\mathcal{L}|$=5 (16PF/Cattell's test with m=163 and EQSQ/Baron-Cohen's test with m=120) to predict the self-rated accuracy ($|C|$=3).

Relevance of Non-Constant Biclusters. Table 2.3 motivates the need for integrating multiple coherencies for real data analysis, measuring its impact on the: percentage of confident decisions (testing observations with over 10 matches and a single class with distinctive higher probability), the average bicluster size, and the weighted lift (Def.VI-1.6) of the discovered rules. For this analysis, FleBiC was parameterized with δ=1/6 coherency strength,

[3] Available in http://web.ist.utl.pt/rmch/software/bclassifier
[4] http://eps.upo.es/bigs/datasets.html
[5] http://eigentaste.berkeley.edu/dataset/
[6] http://personality-testing.info/

10% noise-tolerant checks, merging with 70% overlap, and decreasing support until a minimum number of 50 significant rules per class is found. From this analysis we observe that modeling biclusters tolerant to noise, not susceptible to discretization problems and following flexible coherencies is key to better discriminate classes (+20pp). The gains increase when moving from the isolate use of each coherence towards their integrated use (+10pp) as each coherency models unique local regularities. This improved ability to discriminate classes seems to be also correlated with the larger size of biclusters (each with 5 to 10 new features) and higher correlation strength of decision rules.

Coherence	Percentage of Highly Confident Decisions				Number of Features N(μ,σ)				Average Weighted Confidence			
	Colon	*Lymph*	*Embryo*	*Leukemia*	*Colon*	*Lymph*	*Embryo*	*Leukemia*	*Colon*	*Lymph*	*Embryo*	*Leukemia*
Constant (baseline)	**0.45**	**0.52**	**0.42**	**0.49**	**10±2**	**8±2**	**11±3**	**10±3**	**0.82**	**0.94**	**0.81**	**0.92**
Constant (noisy)	0.69	0.73	0.67	0.72	16±4	13±3	15±4	17±4	0.81	0.94	0.82	0.92
Symmetric	0.66	0.71	0.65	0.68	18±4	14±3	16±4	19±4	0.79	0.91	0.80	0.91
Additive	0.70	0.73	0.68	0.73	24±3	16±3	22±4	26±4	0.79	0.89	0.81	0.91
Multiplicative	0.69	0.71	0.67	0.68	20±3	14±3	19±4	23±4	0.79	0.88	0.80	0.90
Orde-Preserving	0.65	0.71	0.62	0.69	31±5	21±4	29±5	38±5	0.80	0.88	0.80	0.89
Plaid	0.70	0.71	0.69	0.72	22±3	16±2	20±4	24±4	0.80	0.90	0.81	0.90
Integrated (FleBiC)	**0.83**	**0.90**	**0.82**	**0.88**	**21±4**	**14±3**	**18±4**	**22±4**	**0.89**	**0.97**	**0.90**	**0.95**
	JesterD$_1$	*JesterD*$_2$	*16PF*	*EQSQ*	*JesterD*$_1$	*JesterD*$_2$	*16PF*	*EQSQ*	*JesterD*$_1$	*JesterD*$_2$	*16PF*	*EQSQ*
Constant (baseline)	0.41	0.42	0.40	0.39	5±1	6±1	5±1	6±1	0.79	0.77	0.73	0.72
Integrated (FleBiC)	0.58	0.56	0.52	0.50	9±2	11±3	7±2	9±2	0.85	0.85	0.78	0.77

Table 2.3: Impact of learning from non-constant biclusters over real data. Results based on top-100 rules, with 30% to 50% supporting (class-conditional) observations. Criteria: *1)* percentage of testing observations (from 10 cross-fold validation) with >10 matchings and clear preference towards a single class (>10%), *2)* number of features, and *3)* the proposed weighted confidence.

Table 2.4 analyzes some of the properties of the learned regions for both biological and social data contexts. This analysis demonstrates the relevance of using non-constant coherency assumptions with tolerance to noise to find regions of interest, further supporting the observed improvements provided in Table 2.3.

Data domain	*Notes*
Gene expression *(colon, embryo, lymph, leukemia)*	Biclusters define subsets of subjects/samples where a subset of genes show a regulatory profile correlated with the phenotype under classification. Non-constant coherencies are essential to model different expression levels across subjects/samples explained by shifting, scaling or ordering factors due to structural differences on the responsiveness of genes across individuals and biases from the applied measurement and preprocessing.
Collaborative filtering *(Jester D_1/D_2)*	Discriminative biclusters allow the isolation of users with different tastes, as well as subcategories of items (jokes) possibly correlated with the item (joke) with preference under prediction. The scaling, shifting and ordering factors are critical to guarantee an identical variation of preferences, and thus accommodate coherent differences in the rating scale (-10 to 10). Plaid and symmetries are not relevant in this context (their selection does not impact solutions).
Trait surveys *(16PF and EQSQ questionnaires)*	Discriminative biclusters define subsets of individuals with a shared subset of psychophysiological traits possibly correlated with the self-rated accuracy (class). Scaling, shifting and ordering factors are critical to deal with the subjectivity of answers' scales, as some individuals tend to better explore the scale (1 to 5) independently of the assessed traits. Noise is important to accommodate subtler deviations in profiles and possible inaccuracies during the answering.

Table 2.4: Properties of the learned discriminative regions from the selected biological and social data contexts.

Table 2.5 provides details on illustrative discriminative biclusters (with shared pattern φ_B but varying properties such as higher support associated with non-constant coherencies). This analysis shows the importance of using noise-weighted criteria to better model the support of biclusters (and, consequently, to score rules), and the relevance of using flexible coherencies to increase the probability of matches and thus alleviate the common downsides of associative models.

FleBiC's Performance. To complement the previous analyzes, Figure 2.3 assesses the variations in performance over real data (under 10 cross-fold validation) from parameterizing FleBiC with varying coherencies. This analysis

data	φ_B	$\|\varphi_B\|$	*coherence*	*noise*	*abs. support* {C, C\C}	*weighted abs. support*	*Integrative score*	*%test matches*
leukemia	P_1	9	constant	0.0	{9,0}	{10.3,0.9}	0.92	14%
leukemia	P_1	9	constant	0.2	{16,0}	{10.3,0.9}	0.92	14%
leukemia	P_1	9	additive	0.1	{24,0}	{17.4,2.1}	0.89	22%
colon	P_2	8	constant	0.0	{14,0}	{16.2,1.9}	0.89	11%
colon	P_2	8	const. plaid	0.1	{19,1}	{15.8,1.6}	0.91	13%
colon	P_2	8	multiplicative	0.1	{20,1}	{17.9,2.4}	0.88	15%
jester D_1	P_3	7	constant	0.2	{162,6}	{311.2,22.9}	0.81	9%
jester D_1	P_3	7	order	0.1	{421,9}	{602.7,61.6}	0.90	17%

Table 2.5: Illustrative rules with fixed pattern φ_B (P_1={6,5.4,2.7,8.4, -3.6,6.3,5.1,8.1,7.5}, P_2={-3.3,-5.4,-5.7,-6,-2.7,-7.8,-8.1,-6.3}, P_3={-6.8,-6.7,-3.2,3.3,3.6,6.9,7.2} when $a_{ij}\in$[-10,10]) and varying coherence. Comparing their (weighted) absolute support per class, integrated score and percentage of testing observations (under 10-CV) with θ=0.8 matching threshold.

show significant improvements (t-tests with p-value<0.05) in terms of accuracy and sensitivity[7] from the integrated use of different coherencies, possibly explained by the focus on new biclusters and the less strict matches associated with flexible coherencies.

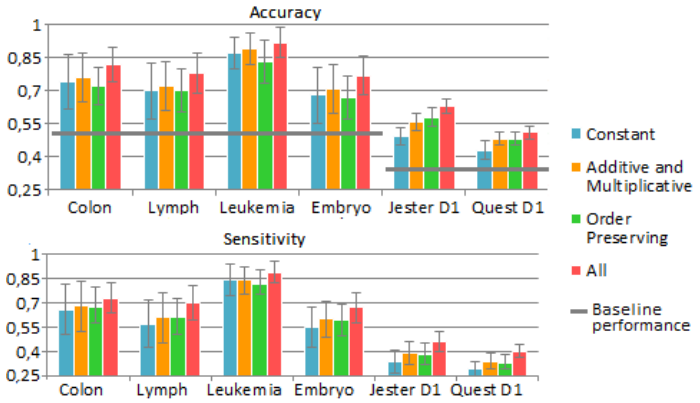


Figure 2.3: Accuracy and sensitivity gains of modeling non-constant biclusters from high-dimensional biomedical data.

Finally, Figures 2.4 and 2.5 extend the analyzes provided in Figures 1.4 and 1.5 to validate whether FleBiC is able to accurately and efficiently perform classification in synthetic datasets with planted regions following non-constant coherencies (properties according to Table II-3.4). FleBiC's performance is compared against classifiers based on discriminative pattern mining (using CMAR [401] after data discretization accordingly to δ) and discriminative biclustering (using FDCluster [661]), and global classifiers based on support vector machines (SVM) and Bayesian networks (BayesNet) from Weka [286].

Figure 2.4 assesses FleBiC's ability to correctly classify observations based on planted regions with varying coherence strength, number of supporting observations, amount of noise, and discriminative power. The gathered results confirm the distinct superiority of FleBiC against both associative and global classifiers when considering either constant or non-constant regions, motivating the importance of discovering flexible coherencies robust to noise and of using adequate scoring criteria.

Figure 2.5 assesses the performance of FleBiC for varying data contexts. Results confirm its competitive, and sometimes superior accuracy, showing its critical ability to select regions with flexible coherence and quality (no-propensity to discretization problems). The levels of efficiency in Figure 2.5 show that, although FleBiC's efficiency is penalized by the increased complexity associated with the discovery of biclusters with flexible coherence, it is able to deal with high-dimensional data.

[7] The case class in biological data and the less accurate class in social data.

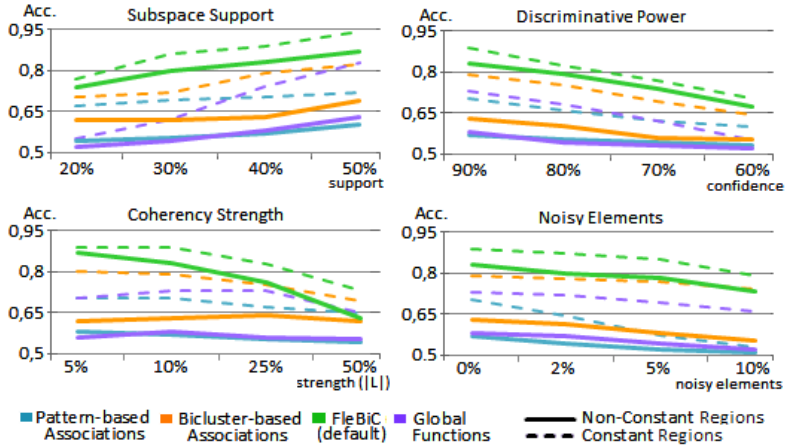


Figure 2.4: FleBiC's ability to learn from discriminative biclusters against peer classifiers in the presence of regions with varying support, discriminative power, coherency strength, coherency assumption and noise.

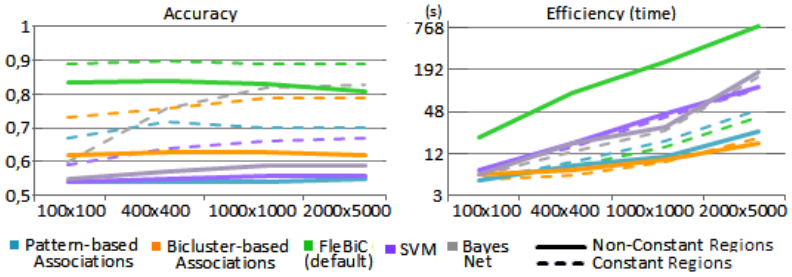


Figure 2.5: Accuracy and efficiency of FleBiC against state-of-the-art classifiers over synthetic data.

2.4 Summary of Contributions and Implications

This chapter extended the previously considered scope of associative classifiers towards data contexts characterized by the presence of non-trivial yet meaningful and coherent regions. In this context, we motivated and quantified the impact of discovering discriminative biclusters with non-constant coherencies for classification tasks.

For this end, we extended BiC along its three major steps. First, we provide an integrative discovery of discriminative biclusters with additive, multiplicative, symmetric, order-preserving and plaid assumptions and varying quality, guaranteeing: 1) an adequate space exploration and 2) combination of regions with distinct properties by considering neither too restrictive nor loose forms of dissimilarity. Second, we extend BiC's scoring schema to correctly weight biclusters with distinct coherency assumptions in order to avoid that the learning is dominated by a subset of relevant regions. The quality of biclusters is also included in the integrative score to guarantee a preference towards not only large but also noise-intolerant regions. Finally, the testing phase is enlarged with new criteria to adequately match observations against non-constant biclusters.

Results gathered from both real and synthetic (high-dimensional) data confirm the underlying hypothesis of our work: modeling regions with non-constant coherencies and varying quality improves the performance of associative classifiers. Results also confirm the superiority of FleBiC against state-of-the-art classifiers.

Future Work. Three major directions are identified for upcoming research. First, we expect to rely on the principles proposed in the context of BicNET to adequately learn from sparse data. Second, since real data is commonly described as a mixture of both local and global regularities, we aim to study the synergies between associative and global classification models. Finally, we expect to extend FleBiC to not only accommodate the target coherency assumptions in this chapter, but additional forms of homogeneity given by alternative biclustering merit functions.

3

Advanced Aspects of Associative Classification

Although research on classification from tabular data has been increasingly matured through the last three decades, there are still open challenges and unprecedented opportunities that have been driven an increasingly attention more recently in the research community.

First, the need to learn classifiers from sparse data, sometimes referred as sparse classification. In this context, it is critical to adequately learn from: structurally sparse data (possibly derived from biological and social network data), missing data (where a robust handling/interpretation of missing elements is essential across medical data domains) and uninformative data (data with non-missing yet irrelevant elements that are apriori known is highly common across biomedical and social data domains).

Second, the need to learn from non-trivial tabular data, including data contexts characterized by: 1) the simultaneous presence of local and global regularities, 2) regions without well-defined boundaries (better described by the probability to which a given observation and/or feature belongs to a given region of interest), and by 3) the presence of regions spanning non-identically distributed features.

Finally, the need to effectively incorporate the increasingly available background knowledge (from knowledge repositories, literature and user expectations) to guide the learning task, thus exploring both effectiveness and efficiency gains. Similarly to the previous challenges, this opportunity is important to guarantee an adequate focus on regions from high-dimensional, thus reducing the propensity of classifiers to under/overfit the input data.

This chapter aims to address these challenges and opportunities from the less studied angle of associative classification. As a result, it provides four major contributions. First, a new classifier able to learn from sparse data. Second, principles to compose ensembles of classifiers with contrasting behavior. Third, a discussion on the benefits and limitations of learning stochastic descriptive models for associative classification. Finally, principles to specify and effectively incorporate available background knowledge with positive impact on the learning. Figure 6.1 synthesizes the tackled challenges and contributions.

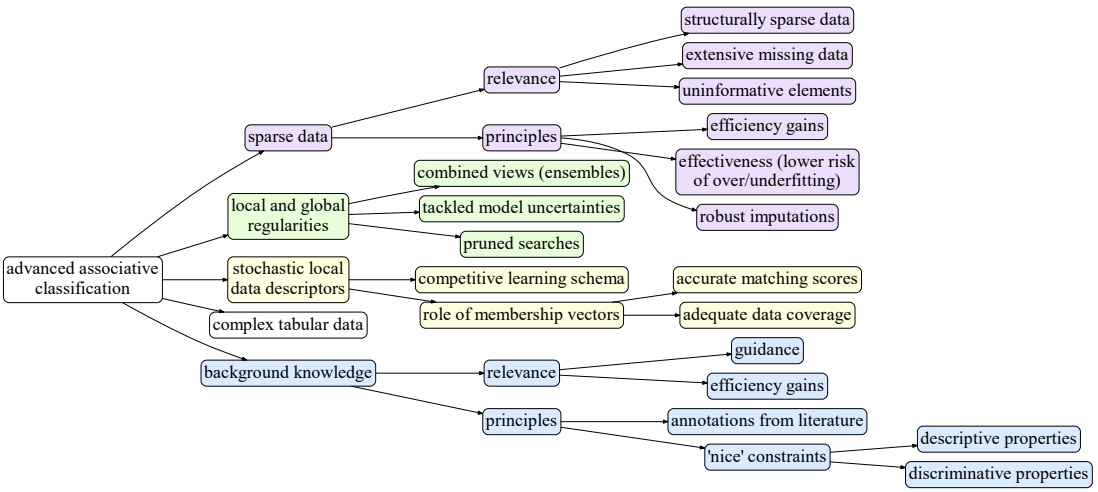


Figure 3.1: Synthesized view on the tackled challenges and seized opportunities to improve associative classification.

This chapter is structured as follows. *Sections 3.1* to *3.5* respectively tackle the problem of learning from: 1) sparse data, 2) data with regularities of varying extent, 3) stochastic descriptors of data, 4) complex tabular data domains, and 5) data domains with available background knowledge. Finally, a summary of the contributions is provided.

3.1 Learning from Sparse Data

Background. Three major factors impact the sparsity level of a given tabular dataset: 1) true missings, 2) false missings, and 3) uninformative elements. First, structurally sparse data is characterized by the presence of true missings, associated with unnecessary or redundant values and with meaningfully zero entries (e.g. disconnected nodes from network data).

Second, sparse data can be complementarily characterized by an arbitrary number of false missings, typically associated with monitoring holes, default expectations, errors, among many other possible reasons for the absence of imputed values. False missings are common among medical data domains as well in biological and social domains where the access to certain data is prevented due to high costs or security reasons.

Finally, sparse data can be additionally associated with the a-priori knowledge of non-interesting elements, which can be seen as candidates for removal. Illustrating, in omic data domains, uninformative elements are typically associated with non-differential regulation of genes or non-differential concentration of molecular entities. For other domains, uninformative elements may correspond to: entries with low-counts in tabular data extracted from text, inconclusive ratings in collaborative filtering data, unprofitable decisions from trading data, or healthy evaluations from medical data. Among these data contexts, the removal of these uninformative elements is desired to guarantee a focus on relevant elements and possibly explore efficiency gains.

In this context, there is the need to learn classifiers in general (and associative classifiers in particular) able to: 1) discard true missings, 2) robustly interpret false missings, and 3) flexibly neglect uninformative elements. However, despite the relevance of this task, classification models from tabular data are not able to deal with sparse data. The large majority relies on simplistic imputation methods to guarantee their applicability. In this context, they show an unnecessary inefficiency and ineffectiveness as they are not able to flexibly remove true missings.

Solution. According to the principles proposed in *Section III-9.2*, BicPAMS is able to learn biclusters from sparse data with an arbitrary-high number of true missings (based on the contributions proposed in *Chapter III-8* and uninformative elements. Since BicPAMS relies on pattern-based biclustering algorithms (which in turn are based on pattern mining searches essentially prepared to mine transactions/sequences of varying length), BicPAMS is able to elegantly exclude data elements from searches. In this way, the searches can focus on regions of interest and explore large efficiency gains both in terms of time and memory.

As a consequence, since FleBiC relies on the class-conditional application of BicPAMS without interfering with its original behavior, it naturally benefits from these underlying advantages. FleBiC is also able to adequately discover discriminative regions with true missings since BicPAMS is additionally able to identify non-dense biclusters with a parameterizable maximum fraction of missing elements. Resulting from previous principles, FleBiC can be termed as a sparse associative classifier.

Furthermore, according to the principles proposed in *Section III-3.2.3*, BicPAMS is also able to adequately handle false missings by providing multi-item imputation methods. In the context of FleBiC, this implies that the imputations are computed from the observations with the same class. As such, FleBiC is also able robust to false missings.

Empirical evidence for the relevance of the enumerated principles to discover regions from sparse data were presented in *Sections III-8.4* and *III-2.7.2*. Moreover, the soundness of FleBiC from these regions was empirically shown throughout the two previous chapters.

The proposed principles are not only applicable to FleBiC but extensible to other associative classifiers from

regions given by discriminative patterns or/and pattern-based biclusters.

3.2 Integrating Associative with Global Functions

Background. Different testing observations may be classified with different degrees of confidence due to the extent of matches and the consistency of labels from the matched rules. In particular, three undesirable situations can occur: *1)* few matches or matched biclusters with low scores, *2)* no label with significantly higher probability (weak consistency of rules' consequent), and *3)* observations not only characterized by local regularities but also by global regularities.

The contributions proposed in the context of FleBiC aim to minimize the first two problems. In particular, we guarantee an adequate: 1) space exploration by focusing on a wide diversity of regions characterized by a varying quality, coherency strength and coherency assumption; 2) scoring schema that guarantees that the learning is not jeopardized by a compact set of regions; and 3) matching criteria with relaxations to guarantee a more informative description of testing observations based on a high number of partial matches. However, the third problem is not yet tackled.

Solution. In order to guarantee that the third problem is adequately addressed we propose the integration of FleBiC with other classifiers. Two major principles are proposed towards this end.

First, FleBiC is extended to combine the (probabilistic) decision outputs of global kernels given by well-known classifiers. By default, FleBiC uses the output of support vector machines, Bayesian networks and multivariate discriminants due to their contrasting behavior against associative classifiers. Given a testing observation, consider $\mathbf{p}$ to be the ouput vector with the (normalized) strength per class, $\mathbf{p}=\{p(c_1|x_{new}), .., p(c_{|\mathcal{L}|}|x_{new})\}$. In this context, FleBiC's output, $\mathbf{p}_L$, is weighted with the output of d classifiers with global functions ($\mathbf{p}_{Gi}$):

$$\mathbf{p} = \alpha \mathbf{p}_L + \frac{(1-\alpha)}{d} \sum_{i=1}^{d} \mathbf{p}_{Gi} \tag{3.1}$$

where, from empirical evidence, $\alpha \approx 0.4$ (default parameterization).

Second, in the presence of matches but not a delineated preference towards a single label, the labels with significantly low probability to occur for a given testing observation can be excluded and not used as input to train the global classifiers. In this context, only the class-conditional partitions associated with the most promising labels are considered

Contrasting, when there is less than $2 \times |C|$ matches, the inverse strategy is considered: the $C \subseteq C$ labels with lower probability from global classifiers are excluded from the $\mathbf{p}_L$ calculus, $p(c \in C|x_{new})=0$. As such, this principle can largely minimize unnecessary biases.

Empirical Evidence. Figure 3.2 quantifies the gains from integrating the output of FleBiC with global classifiers (given by Bayesian networks, support vectors machines and multivariate discriminant functions) over biological and social data (details in *Section VI-1.4*). The observed results support the relevance of the provided principles.

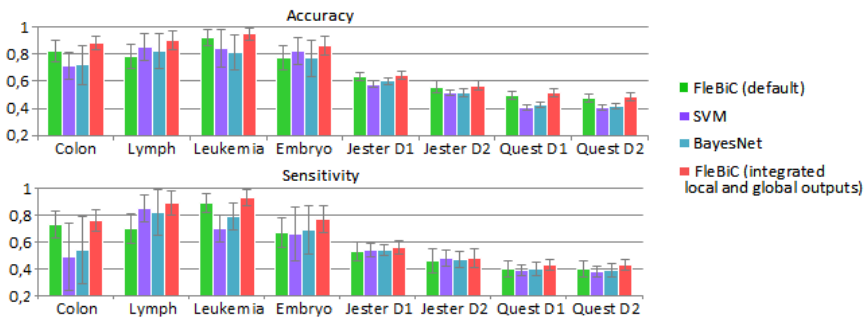


Figure 3.2: Gains in accuracy and sensitivity from integrative local and global classification decisions over real data.

3.3 Stochastic Learning from Generative Biclustering Models

Background. Associative classifiers are deterministic in nature as they rely on deterministic descriptors of data (discriminative patterns and biclusters). However, as largely discussed throughout *Chapters IV-3* and *IV-4*, stochastic descriptors of data can be similarly used to model relevant regions, holding particular properties of interest. In fact, a large number of stochastic biclustering methods have been proposed [324, 449, 573, 113, 527, 581, 72, 584], as well as diverse attempts of defining stochastic methods for pattern mining [629, 347, 659, 6, 66, 392, 389]. However, despite the availability of such methods[1], their relevance for (associative) classification has not yet been target and therefore it is unclear.

Solution. To compare the relevance of deterministic versus stochastic biclustering models for (associative) classification, we first define their properties. According to Def.I-1.10, given a tabular dataset, a biclustering model is given by a set of biclusters, where each bicluster is defined by a subset of observations and features satisfying certain criteria of homogeneity, discriminative power and significance. Deterministic approaches for biclustering either rely on exhaustive or iterative/recursive searches to identify a set of well-defined biclusters (see Table III-1.3). Contrasting, stochastic approaches for biclustering learn a parametric model according to a likelihood function [324, 449]. The biclustering model is either given by: 1) the parametric model (often given by a factorial model defined by a composition of continuous latent variables), or by 2) a set of well-defined biclusters derived from the model's parameters (often recurring to a collapsed Gibbs sampler or peer procedures to infer the posterior distribution of the binary bicluster membership from these models). As such, stochastic biclustering models are, in their purest form (before sampling), typically characterized by a set of vectors defining the probability that a given observation or feature belongs to a given bicluster. Figure VI-3.3 provides an illustrative stochastic biclustering model.

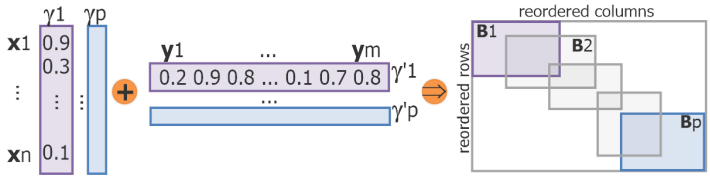


Figure 3.3: Stochastic biclustering models: presence of membership vectors yet inflexible structures and homogeneity.

In this context, learning a classifier from the introduced parametric models naturally differs from associative classifiers inferred from deterministic biclustering models. As such, the learning schema needs to be adequately revised to learn effective classifiers from stochastic biclustering models. *Pointer 3.1* provides a classifier proposed by the authors towards this end. The introduced classifier benefits from the presence of membership vectors. First, it enables more accurate matching scores based on the generative probability of a given testing observation being described by the discriminative stochastic biclusters. Second, it gives the possibility to reduce the cut-off threshold criteria (for deciding whether a feature or observation participates in a given bicluster) in order to guarantee a better coverage of the data space.

Pointers 3.1 Illustrative classifier from stochastic biclustering models

A classifier from stochastic biclusters can be designed by considering variants of the discovery, training and testing steps of the previously proposed classifiers. First, during the region discovery step, $|C|$ sets of class-conditional regions are discovered. Each region is essentially characterized by two vectors of membership probabilities (features and observations) and a third vector with the expected pattern $\varphi_{\mathbf{B}}$ spanning all the probable features inferred from the observed values for the most probable observations. Second, during the training stage, decision rules are inferred by assessing the discriminative power of the retrieved stochastic regions. This is done by assessing whether similar subsets of most probable features (and their patterns) are also discovered for other class-conditional partitions. This can be accomplished by testing the similarity of a given feature membership vector (and respective pattern) against feature membership vectors from other classes. The confidence of a rule is defined by the computed

[1] Stochastic methods for biclustering and pattern mining should not be mingled with the sparse kernels described in *Section 1.2* [78, 215, 378, 214, 217]. Sparse kernels are neither able to flexibly model regions of interest.

metric of similarity. Finally, during the testing stage, the class-conditional fit of an unlabeled observation is assessed against the feature membership vector and pattern of all discriminative biclusters. The feature probabilities are used to uniformly weight the matching score.

As a result, classifiers from stochastic descriptive models hold the benefit of: 1) minimizing the generalization error, thus reducing their underfitting propensity for high-dimensional data, and 2) address the problem related with the possible inaccuracy and scarcity of matches between a test observation and the learned regions.

Despite the relevance of these two properties, stochastic biclustering models suffer from additional downsides. First, stochastic methods for biclustering place restrictive constraints on the allowed number, positioning, coherency and quality of the biclustering model. This prevents the identification of a flexible structures of relevant regions with varying coherency and parameterizable quality. Second, these methods are easily prone towards efficiency bottlenecks in large-scale data contexts (see results provided in *Section III-8.4*). Finally, and contrasting with classifiers based on pattern-based biclusters, the use of stochastic biclustering models does not support learning from sparse data, multi-value assignments, an effective incorporation of background knowledge, among other benefits of the previously proposed associative classifiers.

3.4 Learning from Complex Tabular Data

Background. Although many of real-world labeled datasets have non-identically distributed features, associative classifiers inferred from discriminative biclusters are not able to adequately learn from these data contexts. This is due to the fact that the majority of discriminative biclustering methods are only prepared to learn from data with identically distributed features (see *Section III-3.1.3*). Complementarily, since associative classifiers reliant on discriminative patterns are only prepared to interpret nominal features, they are not able to adequately learn from data with (discretized) numeric features and ordinal features. In this context, enhancing the proposed associative classifiers is essential to adequately learn from mixtures of (non-identically distributed) nominal, ordinal and numeric features, as commonly observed in clinical and heterogeneous data (see *Chapter III-3*).

Solution. According to the principles proposed in *Section III-3.2.3*, BicPAMS is able to learn biclusters from complex tabular data. Furthermore, when BicPAMS is parameterized with the statistical principles proposed in *Section V-3.2*, it is additionally able to provide guarantees on the significance of these biclusters.

According to the behavior described in *Section VI-1.3.1*, FleBiC relies on class-conditional BicPAMS searches to compose decision rules. Since the behavior of FleBiC does not interferes with the discovered regions, FleBiC can soundly rely on regions from complex tabular data. In particular, in data domains characterized by the presence of numeric/ordinal features and nominal features, the homogeneity of the underlying regions can be given by non-trivial yet meaningful mixtures of coherencies, such as illustrated in *Basics III-3.2*.

3.5 Effective Incorporation of Constraints

Background. A largely researched problem in the field of classification is centered on how to incorporate the increasingly available background knowledge to guide the learning process [534, 387, 590]. Despite the substantial evidence for the relevance of using background knowledge to explore efficiency gains and guarantee accurate decisions [590, 534, 84, 264, 158, 387], there are not yet contributions on how this knowledge can be used to adequately affect associative classification. Relevantly, pattern mining has been extended in multiple ways in the context of domain-driven pattern mining for the accommodation of constraints derived from background knowledge. Furthermore, previous work by the authors (*Chapter III-10*) [314] propose an integrative view of domain-driven pattern mining and (pattern-based) biclustering, enabling potential synergies to enhance associative classifiers based on discriminative biclustering models.

Solution. BicPAMS was extended in *Chapter III-10* with principles from domain-driven pattern mining to explore efficiency gains and benefit from the guidance of available background knowledge. In this context, we further showed how functional annotations and constraints with succinct, (anti-)monotone and convertible properties could be effectively incorporated during the biclustering task. In particular, to support constraints on biclusters with different coherency assumptions, we adapted the underlying biclustering searches to support the constraint-based mining of frequent itemsets, association rules and sequential patterns by extending F2G and IndexSpan searches (*Section III-10.4*).

Since FleBiC relies on BicPAMS searches without the need to affect their mining step, it can be used in the presence of the different forms of background knowledge discussed in *Chapter III-10*. In particular, FleBiC can be applied in the presence of an arbitrary number of annotations per observation derived from knowledge-based repositories and literature. This possibility is particularly interesting as these annotations can provide further discriminative criteria to support the classification task.

Furthermore, the exhaustive list of meaningful constraints with nice properties for biological and social data provided in *Section III-10.3.2* can be also considered to either guarantee a focus on specific (possibly non-trivial) regions of interest and to explore efficiency gains from their succinct, (anti-)monotone and convertible properties.

Finally, a new set of constraints with nice properties, yet targeting discriminative aspects of data, can be additionally specified and easily supported. To illustrate some of these constraints, consider a gene expression dataset where samples are labeled with one of the three following classes {cancerA,cancerB,noCancer}. In this context, the user may be interested in discovering regions with specific profiles of interest for each type of cancer, such as regions with a certain gene activated for cancer type-A (class-conditional succinct constraint) and regions with more accentuated repressed expression for cancer type-B (class-conditional convertible constraint).

Alternatively, the user may specify conditions on the desirable levels of discriminative power. This can be easily accomplished through the parameterization of the maximum and minimum support thresholds (either absolute or relative) across classes. Illustrating, the user may be uniquely interested in discriminative regions with a relative weighted support (Def.VI-1.6) above 0.5 for a subset of classes and below 0.1 for the remaining classes. An exhaustive description of the possible types of constraints, as well as their illustrative instantiation across different data domains, is however out of the scope of this work.

3.6 Summary of Contributions

This chapter covered advanced aspects of associative classification that deserve a closer attention. First, we enhance associative classifiers with principles for an adequate learning from sparse data, able to explore significant efficiency gains and robustly deal with missing elements. Second, we show how the benefits from global and local classifiers can be effectively combined through ensemble models to learn from data characterized by regularities of varying extent. In particular, we show how the learning schema and outputs can be adequately integrated to guarantee that loosely discriminative class-conditional models (during training step) or low-confident decisions (during the testing step) can be complemented with information from alternative classifiers. Third, we discuss the benefits and challenges from considering stochastic descriptive models for classification, opening new whole set of directions for future work. Fourth, we show the compliance of FleBiC to learn from tabular data with mixtures of features with distinct domains, a necessary condition to guarantee the proper applicability of the proposed associative classifiers on a wider-set of data. Finally, we show that FleBiC is compliant with the constraints discussed in *Chapter III-10* and can be additionally extended to effectively incorporate expectations on the discriminative aspects of the underlying regions of interest.

4

Learning Associative Classifiers from Structured Data

This chapter extends the scope of learning towards labeled structured data contexts. Learning from labeled structured data is of heightened importance for a wide-range of biomedical and social applications, including: phenotype discrimination from gene expression time series; disease prediction from repositories of clinical events; classification of user behavior from collections of temporal snapshots of a social network; diagnosis from (multivariate) physiological signals; financial decision support from repositories of trading actions; marketing initiatives from (e-)commerce events, and web content organization from user actions. Despite its relevance, the state-of-the-art on learning classifiers from structured data is challenged by the high-dimensionality of these data contexts and by the need to model both structural and temporal relations.

An interesting property associated with the listed data domains is that, under the proposed principles in *Book IV*, they can be mapped into a set of labeled observations, where an observation is either given by a multi-set of events or by a multivariate time series. In this context, the requirements associated with the supervised learning from these data structures can be better specified. The task of learning from labeled multi-sets of events is challenged by two major requirements: 1) modeling temporal dependencies among discriminative events with arbitrary levels of sparsity; and 2) modeling arrangements of discriminative events with cross-attribute dependencies. The task of learning from three-way time series is challenged by the need to discover discriminative cascades characterized by an inherent: 1) a flexible composition of modules capturing coherent changes of values over time, and 2) misalignments between observations.

Despite the numerosity of contributions in the field of classification from structured data, they are not able to adequately answer these tasks. First, classifiers proposed to learn from sequences of events are neither well-prepared to deal with arbitrary levels of sparsity between events nor able to learn from distinct event types (multiple attributes). In this context, principles from integrative learning have been proposed, such as the learning from each attribute separately followed by a voting stage. However, they fail to model important discriminative regularities that only appear when the multiple attributes are analyzed in a combined manner. Second, although a large number of classifiers have been proposed for the analysis of multivariate time series, they are not prepared to model relevant regions given by discriminative cascades. In fact, they are not able to flexibly discard uninformative regions, which is a critical requirement when dealing with time series with a high multivariate order (in biological contexts m>1000).

This chapter tackles the research question of whether it is possible to learn effective (associative) classifiers able to address the criticisms associated with these two learning tasks. Accordingly, Figure 4.1 provides a structured view on the applicability, requirements and related work on these tasks.

For this aim, although these learning tasks appear to be challenged by different requirements, this chapter proposes a consistent view on how they can be similarly answered by mapping them into integrative and labeled time-enriched itemset sequences and by targeting the task of adequately modeling informative and discriminative regions from these data structures to learn effective classification models.

This task is tackled from two angles. First, we propose deterministic associative classifiers able to learn from

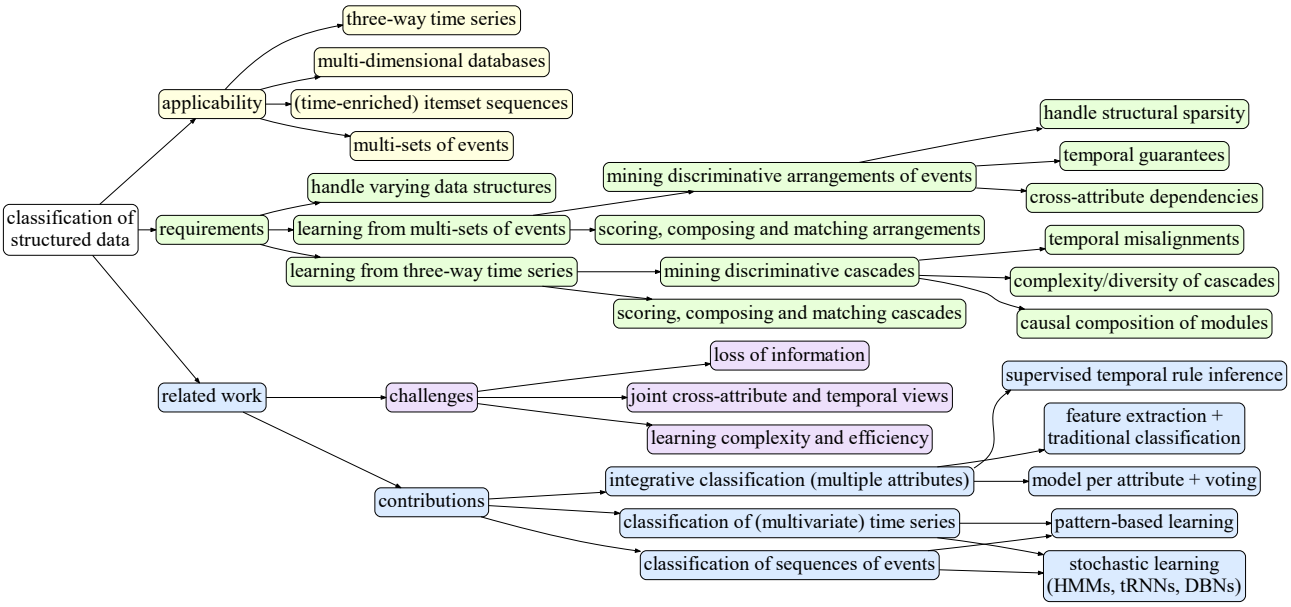


Figure 4.1: Applications, requirements and current challenges to learn classifiers from labeled structured data.

time-enriched itemset sequences based on relevant regions given by discriminative temporal patterns. Second, we propose stochastic classifiers able to learn from itemset sequences based on generative models focused on local regularities. For these ends, we reuse the contributions for the deterministic/stochastic learning of (class-conditional) local descriptive models proposed in the context of *Book IV*, and extend them with adequate discriminative criteria and new training and testing functions. Finally, we refine these contributions to address the specific challenges of learning from multi-sets of events or three-way time series.

As a result, we should be able to answer the following research question: to which extent are the proposed deterministic and stochastic classifiers able to learn from structured data? Figure 4.2 provides a structured view on the solution space. Accordingly, this chapter provides the following major contributions:

- structured view on the contributions and limitations towards the classification of complex temporal data;
- mappings to handle varying labeled data structures (multi-dimensional, relational, multi-sets of events, three-way time series) into an integrative and temporal data structure (time-enriched itemset sequences) conducive to learning;
- new associative classifier based on discriminative, integrative and temporal patterns discovered from time-enriched itemset sequences. In particular, six major contributions:
  - revised notions of support and lift to guarantee sensitivity to noise and temporal misalignments;
  - extended discovery of patterns driven by discriminative power and efficient composition of rules with disjunctions of labels in the consequent;
  - new integrative score based on the (weighted) support, length and lift of temporal patterns;
  - variations on behavior to prioritize occurrences on certain time periods (e.g. favoring of recent events);
  - enriched class strength calculus and new matching criteria for testing observations (sensitive to both structural and temporal misalignments) with degree of relaxations dependent on the number and score of matched regions;
  - specialization of the proposed behavior to adequately model cascades and arrangements of events;
- new stochastic classifiers based on the extension of the proposed (unsupervised) HMMs for the classification of time-enriched itemset sequences, and customization of their properties to adequately model the specificities associated with three-way time series and multi-sets of events;
- systematic comparison of the properties of stochastic learners against deterministic learners;

- principles to handle data structures with a mixture of both temporal and static attributes.

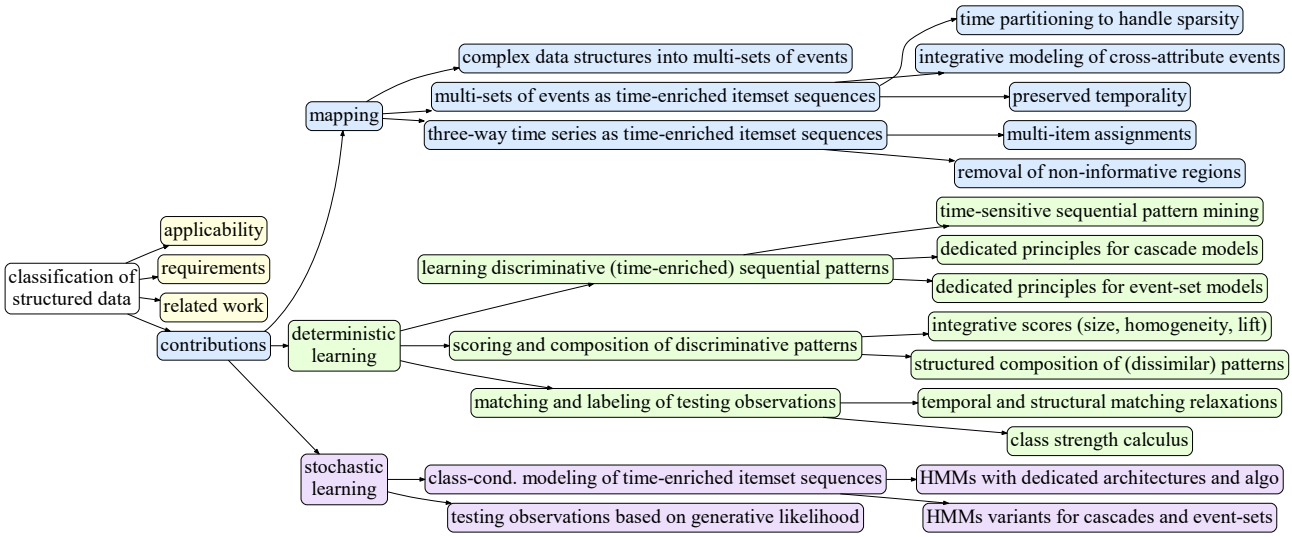


Figure 4.2: Proposed contributions for learning classification models from labeled structured data.

These contributions are integrated within two associative classifiers: a deterministic classifier, P2MID (Pattern-based Predictive Models from Integrated Data), and a stochastic classifier, MaCID (Markov-based Classifier from Integrative Data). Experimental results hold evidence for the utility of the proposed contributions, for the superior performance of P2MID and MaCID against classic classifiers (applied on denormalized data contexts), and their critical role for the analysis of structured data contexts with non-trivial mixtures of (possibly temporal) attributes.

This chapter is organized as follows. *Section 4.1* provides the background notions on learning classifiers from structured data. *Section 4.2* extensively surveys contributions from related streams of research to learn from temporal data with multiple attributes. *Section 4.3* describes the solution space: proposes associative classifiers – P2MID and MaCID – learned from integrative temporal data. *Section 4.4* compares and discusses the performance of these classifiers against classic classifiers over real data. Finally, the contributions and implications are summarized.

4.1 Background

The structured data contexts target in this thesis can be essentially defined by a multiplicity of temporal attributes, where each attribute is characterized by a set of occurring values in time. In this context, an attribute can be described by a single value, an ordered set of values (sequence), an ordered set of values equally spaced in time (univariate time series) or by a set of timestamped values (event-set). When considering the multiplicity of attributes, observations became described by three-way time series (assuming attributes have same domains and length of occurrences), multi-sets of events or even less-trivial combinations of attributes with dissimilar domains. Let us consider the illustrative case of learning from repositories of health-records to discriminate a medical condition. These repositories can be mapped into structured data domains characterized by n observations (patients) and the presence of a high-multiplicity of m attributes related with: 1) [event-sets] diagnoses, treatments, prescriptions and lab records, 2) [single values] genetic markers and the patient profile, and 3) [sequences] feature vectors capturing the progression of the health state of the patient.

Learning from these data contexts is hampered by challenging requirements, whose definition in the context of unlabeled observations was already provided throughout *Book IV*. *Section IV-1.1* systematized the requirements for learning local models from three-way time series. In this context, a region is given by a cascade due to its inherent ability to frame stochastic temporal dependencies associated with the causal elicitation of modules. *Section IV-2.1* discussed the major requirements for learning local models from multi-sets of events. In this context, a region is

given by an informative arrangement of temporally-related events from (possibly) distinct attributes.

When moving from unlabeled to labeled contexts, three new *requirements* need to be satisfied. First, the modeled regions must be discriminative in order to guarantee their relevance for classification. Second, these regions need to be adequately scored and composed to produce classification models well-prepared to effectively and efficiently test new observations. Finally, these regions need to be adequately matched against a new/testing observation, and their ability to describe a new observation properly weighted to place decisions.

Def. 4.1 Let a structured data space $\mathcal{X}$ be characterized by a set of (temporal) attributes $\{y_1, .., y_m\}$ with structured domains $\mathcal{Y}_i$. Given a set of observations $\{\mathbf{x}_1, .., \mathbf{x}_n\}$ with values in $\mathcal{X}$ and labeled with a class in C, the *classification* task is to learn a model $M : \mathcal{X} \rightarrow C$ to label a new observation $\mathbf{x}$, $c = M(\mathbf{x})$.

Applicability. The presence of labeled structured data provides an unprecedented opportunity to support biomedical, administrative and social decisions. Illustrative applications across these domains include: computer aided decisions for disease diagnosis/prognosis and treatment care (personalized medicine) from multi-sets of events derived from repositories health records; modeling of discriminative regulatory responses in the context of a specific drug, stimulus or phenotype; classification of user profile or behavior from the user actions and social interactions; rating of specific objects based on temporally-enriched collaborative filtering data; suggestions based on web navigation and commercial activity; support to financial decisions based on the market transactions.

4.2 Related Work

Learning from temporal data has been largely researched, with multiple categorization attempts, covering different learning tasks and data structures [559, 128, 191]. *Section 4.2.1* overviews state-of-the-art deterministic and stochastic classifiers from temporal data structures. Due to the inherent difficulty of most of these contributions deal with multiplicity of attributes, *Section 4.2.2* surveys integrative approaches for classification towards this end.

4.2.1 Deterministic and Stochastic Learning of Classifiers from Temporal Data

Classification from temporal data has been mainly driven by the task of labeling unlabeled sequences based on a training set of labeled sequences. Sequences are here generically referred to either consider (numeric or symbolic) time series or orderings of events. In this context, three major types of classifiers can be identified: distance-based, pattern-based and generative classifiers. Given a similarity metric, a target sequence can be labeled by a *distance-based classifier* based on the observed labels from the closest sequences [677].

A *pattern-based classifier* learns a model based on a set of prototype features for each class, and labels a new sequence based on the class with the closest features [535]. When temporality is mapped into phasic features, the model combining these features can be given by traditional classifiers, such as decision trees, Gaussian mixtures, support vector machines (SVMs) or neural networks (NNs) [1, 535, 107, 482]. Extended classification models have been proposed in the presence of sliding features extracted from different sequence segments [362]. Alternative features with impact on the learned models include: sequential patterns to focus on discriminative precedences among events [635], motifs to focus on discriminative sets of events recurring along the sequence [490], wavelets [562], alignment-free dictionaries to guide the learning when specific occurrences are annotated [332], among others [466]. Koch and Naito [371] proposed principles for the extraction of features from sequences with arbitrary-high multivariate order.

A *stochastic classifier* generally learns one generative model for each class-conditional set of sequences, and tests a new sequence based on the class associated with the generative models that describes the testing sequence with highest likelihood. Common stochastic methods for the classification of sequences include formal languages [390, 241], hidden Markov models (HMMs) [74, 478], dynamic Bayesian networks (DBNs) [478, 250] and time-

sensitive Neural Networks (NNs) [369, 652, 433]. Different architectures and algorithms with impact on the learning behavior of stochastic methods have been compared [37, 269], as well as diverse principles to learn from multivariate sequences [368, 478].

Three major limitations are associated with the surveyed classifiers. First, they are not able to deal with sequences with non-fixed multivariate order. This is a necessary condition for the analysis of multi-sets events characterized by an arbitrary number of co-occurrences per time partition, and a desirable condition for the analysis of three-way time series since it enables the possibility of removing uninformative elements and assigning or imputing multi-items per element. Second, these contributions are insufficient to deal with the structural sparsity of multi-sets of events and are associated with the loss of temporal distances between events (generally focused on the observed orderings only). Finally, existing contributions from multivariate time series analysis and multivariate responses prediction are not applicable for multi-sets of events since different attributes (event types) can have different domains and a non-fixed number of events per observation.

4.2.2 Learning Integrative Models from Structured Data

Despite the relevance of the surveyed classification models, they are not able to effectively learn from structured data characterized by a multiplicity of (temporal) attributes with possibly different domains. An attribute is either associated with a type of event in the context of a multi-set of events or with a univariate parcel of a multivariate time series. A naïve solution to learn from these data contexts is to fix the number of occurrences per attribute and mapped them as features (static attributes), followed by the application of classic classifiers. Understandably, this option is only of interest for data with compact number occurrences per attribute or where labels are determined by the most recent events. To tackle this problem, integrative classification models have been recently proposed to deal with the multiplicity of temporal attributes. Table 4.1 surveys such learning settings that can be considered to surpass some of the challenges of the previously surveyed approaches.

Approaches	**Limitations**
1. Learning from features extracted separately for each attribute.	Not able to model cross-attribute dependencies; Loss of information.
2. Learning one model separately for each attribute – using discriminative sequential patterns by ignoring temporal distances or sequence classifiers by removing co-occurrences – followed by a voting stage.	Not able to model cross-attribute dependencies; Loss of information.
3. Extracting integrated features using (bi)clustering views.	Suitability largely depends on the used distances to group observations and attributes; Not mature research stream.
4. Supervised inference of temporal rules.	Not scalable when dealing with large datasets; Complexity.
5. Mapping attributes into feature vectors using aggregation functions (preprocessing data by incorporating multiple attributes at several time points).	Background knowledge required; Loss of information.

Table 4.1: Major directions for the integrative learning of classification models from multi-attribute temporal data.

A first option is to extract features separately from each temporal attribute and then apply a classic classifier [32, 414]. An alternative option is to apply one of the surveyed classification models (see previous section) for each attribute, followed by a voting step [633, 139]. The drawback of these solutions is the loss of critical integrated views that do not emerge when each attribute is analyzed separately. A third option is to extract features from multiple time sequences using (bi)clustering methods that rely on edit-distance metrics based on insert-delete-replace operations [349, 121]. The drawback here relies on: *1)* the complexity of defining effective distance metrics, and on *2)* the suitability of the chosen metrics across observations and attributes. An alternative strategy to avoid sparsity that also results in a significant loss of information is to convert each temporal attribute into a set of time series (also referred as feature vector) by using an aggregation criterion, such as the counting of occurrences across sequent periods or alternative ranking and mean functions [47]. Although the supervised inference of temporal rules can be considered to minimize this problem, this option is not scalable [467]. A final option is to use a preprocessing stage that incorporates multiple attributes at several time points or intervals [47].

However, this solution is only practical in the presence of background knowledge, which is used to select specific occurrences of interest from each attribute.

Figure 4.3 provides an illustrative view on the behavior of these five learning strategies based on the properties of the input data structure. This state-of-the-art analysis clearly highlights the need for more effective classifiers, able to simultaneously model temporal and cross-attribute dependencies.

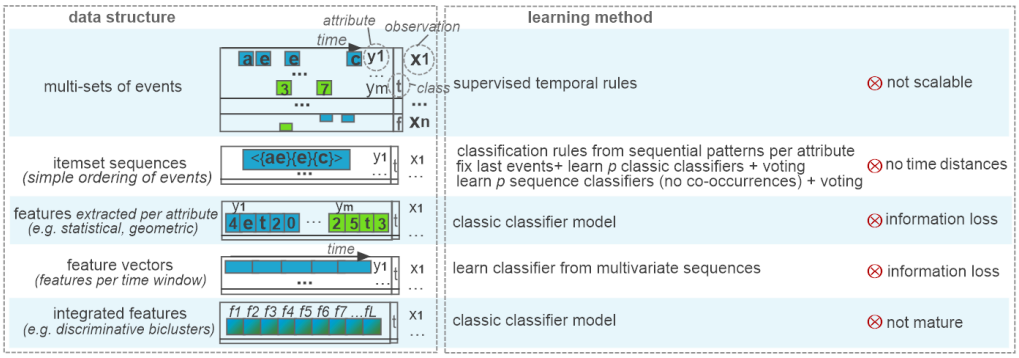


Figure 4.3: Existing data structures and learning methods to model multiple temporal attributes.

4.3 Solution

The solution space for the target learning task is incrementally provided along three sections. *Section 4.3.1* provides the necessary mappings to deal with different data structures and guarantee a standardized and integrative representation of data with multiple temporal attributes that it is conducive to the learning of classification models. Under this mapping, *Sections 4.3.2* and *4.3.3* propose classifiers able to model regions of interest given by integrative temporal patterns using, respectively, deterministic and stochastic learning functions.

4.3.1 Data Mappings to Learn from Structured Data

Multi-sets events. We reuse the mapping proposed in *Section IV-2.3.1* to compose time-enriched itemset sequences from collections of events with multiple types of events (attributes). We reconsider this mapping for classification tasks since: 1) is more conducive to learning tasks (standardized structure offering cross-attribute views and preserving approximate temporal distances), 2) offers an effective way to deal with arbitrary sparsity (use of varying temporal granularities), and 3) enables the use of the proposed methods in *Book IV* for the discriminative and stochastic learning of arrangements of events. This mapping is a result of four major steps: 1) discretization of numeric events (using lengthy alphabets); 2) balancing of the cardinality of the attributes' domain; 3) selection of a temporal granularity and partitioning of the timeline; and 4) grouping of events per partition, removal of duplicates and preservation of empty itemsets. This mapping remains valid in the context of labeled multi-sets of events (labels are preserved during the process). However, since observations with different labels often show radically distinct regularities, both the discretization of attributes and the harmonization of their domains can be applied separately for each class-conditional partition.

Multivariate time series. Despite the large availability of classifiers to model labeled multivariate time series (see previous section), they suffer from two major drawbacks: 1) they are generally not able to focus on relevant regions (such as cascades), and 2) they cannot learn from sparse time series (where uninformative elements are removed) or from time series with multiple items assigned to some of its occurrences. As such, in order to reuse the efforts developed in the context of cascade learning, we propose the mapping of multivariate time series into (time-enriched) itemset sequences and the posterior application of the proposed associative and stochastic classifiers. This mapping (described in *Section IV-1.3*) is a result of three steps: 1) adequate discretization (with estimated coherency strength to produce cut-off points from a dynamically fitted distribution), 2) multi-item assignments

based on distances to cut-off points and removal of elements, and 3) application of transformations (according Def.IV-2.4).

In the context of labeled data, class-conditional differences can be further accentuated to facilitate the extraction of discriminative regions. Illustrating, the observed mean and standard deviation of the class-conditional values per feature or time point can be used to select normalization procedures aiming to heightened the differences between observations with different labels.

Multi-dimensional and relational data. Multi-dimensional data can be mapped as multi-sets of events by deriving events from the entries of a central fact table, grouping these events into a set of observations according to the identifiers of a linked dimension (the split dimension) and with timestamps derived from the date dimension. Relational data can be mapped into multi-dimensional data by selecting the entity-relationship table containing the records of interest and by fixing it as a central fact (linked to additional tables fixed as dimensions). *Section IV-2.3.1* details the properties of these mappings and provided strategies to: 1) surpass memory bottlenecks associated with the high-dimensionality of measures associated with a fact table, and 2) derive events from structured measures.

Labels in these data contexts are commonly derived from one of two sources. First, an additional field from the split dimension (besides its identifier). Illustrating, in the context of healthcare multi-dimensional data, the split dimension is often associated with the patient (or provider) dimension, which may contain relevant fields (including health state, genomic markers, risk profile and psychophysiological traits) that can be used to derive labels. Second, the occurrence of specific events derived from the fact table are critical alternatives to label data. In the context of healthcare data, diagnostics and treatments can be used as the desirable classes/medical conditions.

Mixtures of distinct attributes. Although a large variety of data domains can already be covered by the previously introduced data structures, the proposed mappings are insufficient to learn from data structures characterized by the presence of static and temporal attributes. Illustrating, healthcare data domains can be given by: timestamped events associated with diagnostics and treatments; time series associated with recurrent evaluations; and static attributes capturing the patient profile. In this context, we propose the mapping of these complex data structures into multi-sets of events (which can be consequently mapped as time-enriched itemset sequences).

For this aim, two principles are proposed. First, time series are consistently seen as events equally spaced in time. Second, we consider a static attribute as an event with a special timestamp. Special timestamps are dedicatedly interpreted by algorithms according to their behavior. One option is to use static attributes across observations with a unique timestamp to guarantee their coherent appearance within the target arrangements. A complementary option is to assign a recent timestamp whenever the algorithm benefits more recent events (in order to minimize their exclusion from arrangements). Under these two principles, mixtures of attributes can be soundly seen as multi-sets of events.

Concluding Note. Throughout this section, we extended previously proposed mappings towards labeled data contexts and tackled the yet unanswered problem of dealing with domains characterized by a mixture of distinct (possibly temporal) attributes. A consistent temporal structure resulted from all the previous procedures: *integrative, temporally-enriched and labeled itemset sequences*. For simplicity sake, the following sections assume that this is the observed input data structure, independently of whether raw data is described by multi-dimensional databases, multi-sets of events, three-way time series or any other data structure.

4.3.2 Pattern-based Classifiers for Labeled Structured Data

Under the proposed mappings in previous section, the developed deterministic methods to learn from (temporally-enriched) itemset sequences in *Chapters IV-1* and *IV-2* are in this section extended with: 1) discriminative criteria, 2) scoring and composition schema, and 3) matching and labeling criteria.

Different strategies have been proposed in literature on how to use temporal patterns from itemset sequences

for classification [204, 481]. However, they are only prepared to capture frequent precedences and co-occurrences, and are thus not able to consider temporal distances between items, which is a critical requirement for the target classification models. Additionally, they have been developed in the context of specific data domains and, consequently, the argued levels of performance no longer remain valid for alternative domains.

In order to address these observations, we provide a new associative classifier, referred as P2MID (Pattern-based Predictive Models from Integrated Data), able to model, score and compose discriminative temporal patterns from time-enriched itemset sequences (*Sections 4.3.2.1* and *4.3.2.2*) and test observations against them (*Section 4.3.2.3*). Moreover, the resulting classifier is further enhanced in *Section 4.3.2.4* to guarantee a tuned behavior towards the peculiarities associated with the learning of arrangements of discriminative events from multi-sets of events and discriminative cascade models from three-way time series.

4.3.2.1 Discriminative Temporal Patterns

In its first stage, P2MID generates a set of discriminative time-enriched sequential patterns for each label. According to Def.IV-2.4, a time-enriched sequential pattern is a sequence of itemsets together with their expected time partition of occurrence, P=<$(I_1,\hat{\varphi}_1),..,(I_s,\hat{\varphi}_s)$> where φ_t=median$(\cup_{\mathbf{x}_i\in\Phi_P}\varphi_{k(I_t,i)})$. P2MID computes these temporal patterns by fixing multiple temporal aggregations ($\delta \in \{1,2,..\}$) followed by the discovery of co-occurrences for coarser-grained aggregations under a penalization factor to benefit the discovery patterns that occur for small time intervals. *Section IV-2.3.2* provides further details and illustrations.

P2MID efficiently composes rules with disjunctions of labels in the consequent according to the principles introduced in *Section VI-1.3.1* (joining rules with promising increase of discriminative power and avoidance of redundant calculus by storing relevant counts).

In order to guarantee their discriminative power, P2MID revises the concept of support and lift. First, a new weighted support concept (Def.4.2), that allows for supporting observations to deviate from the expected pattern with regards to: 1) the matched elements (fraction of noisy elements below a parameterizable threshold), and 2) their time frame (allowing adjustable temporal shifts). Second, and grounded on this weighted support concept, we propose a variant of the lift metric to adequately measure the discriminative power of the target rules from temporal patterns (Def.4.2). Similarly to the benefits of using lift enumerated in *Section 1.3.1*, its use in the context of rules from structured data can deal with data imbalance and rules with disjunctions of labels on the consequent.

Def. 4.2 Given a labeled time-enriched itemset sequence $\mathbf{A} \in (\mathcal{X}, C)$, a noise threshold μ, a temporal shift threshold δ, and a decision rule $R: P \Rightarrow C$ in $\mathbf{A}$ (where P is a temporal pattern according to Def.2.4 and C is a subset of labels in C):

- the weighted **support**, sup_P is the number of observations that respect P. An observation $\mathbf{x}$ respects P if it has the at least a fraction of (1-μ) the total items of P occurring on the same time partitions or at least within δ distant-partitions;
- the **lift** of a rule is $lift_R = \frac{sup_R}{sup_P sup_C}$, where: $sup_{R:P\Rightarrow C}$ is the number of observations respecting P (w.r.t. μ and δ) with a class in C, sup_P is the weighted support of P, and sup_C is the fraction of total observations with a class in C.

Illustrating, consider a temporal pattern P=<$(\{a,f\},t_2),(\{d\},t_4)$> and two observations $\mathbf{x}_1$=<$(\{a,f,e\},t_1),(\{g\},t_2),(\{d,c\},t_4)$> and $\mathbf{x}_2$=<$(\{a,b\},t_1),(\{d\},t_3),(\{a\},t_4)$>. Given $\mu=\frac{1}{4}$ and δ=1, $\mathbf{x}_1$ respects P since all the precedences and co-occurrences of P are satisfied ($\{a,f\}$ is observed with one temporal shift and $\{d\}$ occurs in the same exact partition). Since $\mathbf{x}_2$ has $\frac{1}{3} > \mu$ of noisy elements, it does not respect P.

4.3.2.2 Composition of Temporal Patterns

P2MID scores the learned rules using an integrative score according to the support, length, and discriminative power of the retrieved temporal patterns. For this aim, a revised version of (VI-1.2) score is considered. Given an

labeled dataset **A** and a rule $P \Rightarrow C$ in **A**, the proposed integrated score is defined as:

$$\omega_R = \alpha_1 \frac{sup_R}{sup_P}\frac{sup_C}{sup_C} + \alpha_2 \frac{sup_R}{n}\frac{sup_C}{sup_C} + \alpha_3 \frac{\Sigma_{j=1}^{s}|I_j|}{m \times p} \tag{4.1}$$

where m is the number of attributes, p is the average number of occurrences per temporal attribute and $<(I_1, \varphi_1), .., (I_s, \varphi_s)>$ is the temporal pattern. The three parcels of this equation respectively measure the discriminative power, relative support, and relative length of a temporal pattern.

Variations of this score can be easily coded within P2MID. Let us consider the two following variants. First, temporal patterns with short time frames can be preferred in order to avoid the domination of the learning by temporal patterns spanning the whole timeline. Second, temporal patterns with itemsets occurring on certain time partitions can be preferred. Illustrating, more recent occurrences can be prioritized, and thus temporal patterns containing a large fraction of old occurrences can be penalized.

Similarly to CMAR [401], these rules are inserted in a tree structure if: *1)* the χ^2 test over the rule is above a specified α-significance level, and if *2)* the tree does not contain a rule with higher priority (based on the computed integrative score). This tree defines the discriminative pattern-based model, which can be alternatively mapped into simple ordered set of tuples (pattern s, class c, weight β).

Whenever the tree shows imbalance with regards to the number of rules per class and their scores, the tree is pruned based on the computed priorities according to CMAR [401].

4.3.2.3 Testing Observations against Temporal Patterns

Finally, the learned associative model (pruned tree) is used to classify a testing observation by identifying the closest temporal patterns and relying on their matching score for the target labels. The strength of each condition is calculated by computing the WIS score based on all the rules $P \Rightarrow C$ that satisfy a *matching* criterion between the pattern P and the testing observation. The strongest condition, $y \in Y$, is delivered (deterministic output) or, alternatively, the computed strength for each class (probabilistic output). The WIS score is parameterized with the previously defined integrative score: $WIS_c = \sum_{match(R:\varphi_B \Rightarrow C | c \in C)} \frac{sup_c}{sup_C}\omega_R$.

Def. 4.3 Given a rule $R : P \Rightarrow C$ with score ω_R, an observation $\mathbf{x}_{new}$ **matches** P if it respects P with regards to a certain allowed level of noise μ and temporal mismatch δ.

In this context, matching occurs if the the co-occurrences and precedences from a temporal pattern are mostly preserved for the testing observation, as well as their time frame (different time partitions can be observed as long as they preserve the input time shift condition).

The number of shifted partitions is used to penalize the rule score, as well as the level of tolerated noise. According to Def.VI-1.8 both linear and squared penalizations can be considered for this end.

4.3.2.4 Variants to Handle Data Specificities

Learning from Cascades. The proposed associative classifiers can be further extended for the adequate modeling of discriminative cascades given by the learned temporal patterns. First, to guarantee that the target discriminative temporal patterns are robust to noise, the postprocessing procedures described in Table IV-1.1 can be directly applied along the discovery step. These procedures aim to adjust the discovered regions by merging, extending, filtering and reducing sequential patterns. The fact that sequential patterns are now temporally annotated does not interfere with the proposed procedures (essentially based on the similarity between two cascades and of the homogeneity of a single cascade). The time annotations should, nevertheless, be updated whenever a sequential pattern is modified based on the properties of the introduced/removed items and observations. Complementarily, other strategies, such as multi-item assignments, can be considered to increase the tolerance of the target regions

(given by temporal patterns) to noise.

To facilitate the analysis of the learned associative model, the principles proposed in *Section IV-1.3* can be considered to extract modules and their causal or parallel dependencies from a time-enriched sequential pattern. Although this step is optional for classification, it promotes its interpretability.

Finally, the high-dimensionality of data contexts where regulatory/behavioral cascades are observed poses challenges in terms of efficiency. As such the discovery of discriminative temporal patterns should consider some of the principles proposed in *Section IV-1.3.4*, namely removal of uninformative elements, preference towards vertical data formats, searches oriented to model itemsets with a high number of co-occurrences, data partitioning principles, searches for approximative or top-K-dissimilar patterns, and the effective incorporation of constraints associated with preferred forms of homogeneity and expectations on the minimum support, length and items per cascade.

Learning from Arrangements of Events. Contrasting with the learning of discriminative cascades, the use of multiple temporal granularities and of time annotations is mandatory for learning arrangements of events to deal with the arbitrary sparsity of a multi-set of events. Additionally, postprocessing procedures should be also applied in this context to: provide dissimilarity guarantees between arrangements extracted using multiple temporal granularities, and to increase their tolerance to noise. In this context, the similarity between two arrangements of events (for filtering and merging) is given by the alignment-based score in Algorithm 2, and the homogeneity of an arrangement of events (for extensions and reductions) is given by the number of noisy and/or missing events across observations.

4.3.3 Stochastic Classifiers for Structured Data

Contrasting with deterministic classifiers, stochastic classifiers can offer a critical probabilistic and noise-sensitive view of discriminative temporal patterns. To rely on the contributions along *Chapters IV-3* and *IV-4*, we reuse the proposed *hidden Markov models* (HMM) and enhance them to classification. In fact, HMMs have been originally proposed in the context of classification [399], and their maturity, expressive power, inherent simplicity, flexible parameter-control and propensity to deal with temporal data [390] turn them an interesting candidate for the target task. Bilmes [73] refers that there is no theoretical limit to HMMs performance given enough hidden states, rich enough observation distributions, sufficient training data, and appropriate training methods. Furthermore, the limitations[1] associated with the use of HMMs for the unsupervised learning of local descriptive models were addressed throughout *Chapter IV-3*. In this context, we proposed expedite architectures to model co-occurrences and precedences from (temporally-enriched) itemset sequences, data-driven statistics for the adequate initialization of their transitions and emissions, adequate learning settings, and architectures able to model an arbitrary number of (dissimilar) patterns without propensity to spurious background matches. These baseline contributions were extended in *Chapter IV-4* to enable the extraction of temporal guarantees from the learned lattices.

Given the introduced stochastic learning functions, the goal becomes centered on defining adequate training and testing functions for classification. Two options can be considered. The first option is to apply the previous learning setting on each class-conditional set of observations; retrieve informative regions from the learned lattices using the decoding principles described in Algorithm 9; and make use of the associative training and testing functions proposed in the previous section on top of the decoded regions. These regions are given by class-conditional temporal patterns associated with highly probable paths of transitions and emissions. Despite the consistency of this option, it suffers from two major problems. First, since the retrieval of regions is done independently for each class-conditional generative model, there is a heightened propensity that many of the decoded regions are

[1] Major limitations for the application of HMMs to model relevant regions from time-enriched itemset sequences include: 1) restricted task formulations (targeting univariate sequences and contiguous items), 2) the need for prior assumptions (expected number, shape and length of patterns), and 3) inadequate learning settings often characterized by a loose convergence of items associated with spurious background matches.

not discriminative. Second, there is a significant loss of (potentially) relevant information associated with the replacement of the learned class-conditional generative models by the decoded regions.

For this aim, we suggest the use of the default HMM-based classification setting, where the learning is applied on each class-conditional set of observations to produce $|C|$ generative models and then a testing function is directly applied to test the fit of a new observation against each one of the learned class-conditional models. This second option minimizes the previous problems. In particular, the discriminative power is guaranteed during the testing phase by studying the differences in likelihood across the $|C|$ models. Furthermore, since the likelihood of a given model to describe a testing observation is based on the combinatorial sum of the joint probability of all possible paths, no loss of information is incurred.

Since the testing stage can be computationally expensive, we rely on the Viterbi algorithm for its efficient implementation according to the principles for traversing lattices described by Bishop [74].

Variants to Handle Data Specificities. According to the extensions explored in *Chapter IV-4*, the proposed classifiers based on HMMs can be adapted according to the properties of the input data. Illustrating, when stochastically modeling data described by numeric attributes only, continuous HMMs can be considered.

The modeling of cascades from three-way time series can be largely improve through the adequate use of: 1) incremental learning principles (to dynamically adapt architectures throughout the learning based on the observed path convergence), 2) postprocessing procedures, and 3) architectural changes to adequately interpret multi-item assignments. The modeling of arrangements of events is dependent on the: 1) adequate adjustment of the probability associated with emissions of delimeters according to the average sparsity and selected temporal granularity, and 2) effective retrieval of time frames.

The listed contributions are integrated within a new classifier, referred as MaCID (Markov-based Classifier from Integrative Data). MaCID can be further extended to allow the: 1) discovery of discriminative regions of interest by focusing on the differences between class-conditional models, and 2) effective incorporation of constraints based on user expectations by parameterizing the underlying Markov-based architectures.

4.4 Results and Discussion

P2MID and MaCID classifiers were implemented in Java (JVM v1.6.0-24). MaCID was implemented as an extension of the HMM-WEKA package[2]. We run experiments in an Intel Core i5 2.80GHz with 6GB of RAM.

Data. The proposed classifiers were evaluated over the healthcare heritage prize database[3], a structured dataset combining: 1) temporal attributes given by sets of events and feature vectors (time series) collected along two years, and 2) static attributes based on the patient's profile. In particular, it integrates detailed claims (multi-sets of events) and a monthly number of laboratory tests and prescribed drugs (time series) for 113000 patients. According to *Section 4.3.1*, the original relational database was mapped into a multi-dimensional database and finally into a time-enriched itemset sequence for the consistent integration of the different types of attributes.

Three temporal granularities were considered: month, quarter and semester. For the month, quarter and semester granularities, each patient has respectively an average number of 4 (σ=2), 12 (σ=3) and 24 (σ=4) events/items per temporal partition/itemset and 36, 12 and 6 event-sets/itemsets per sequence.

We target the classification of three different medical conditions: *1)* surgery anticipation for the upcoming quarter, *2)* average number of drug prescriptions for the upcoming month (by grouping prescription levels into four classes {0,1,2–5,6+}), and *3)* hospitalization needs in the upcoming period. Upcoming predictions are accomplished by removing the last temporal partition from training data. In particular, for the last task, we also removed health records (HRs) related with hospitalizations, meaning that these predictive models were only learned from

[2]Software available in http://web.ist.utl.pt/rmch/research/software and implemented according to [74, 478].
[3]http://www.heritagehealthprize.com/c/hhp/data (under a granted permission)

attributes related with claims, diagnoses, lab analysis and drug prescriptions. A sample reduction method was applied over the original population to balance the number of observations per class. A random population of 20000 patients with heightened clinical activity was consider for each one these three tasks.

Evaluation Setting. P2MID and MaCID were compared against classic classifiers (average of C4.5, kNN, Naive Bayes, NN and SVM from Weka [287]) by mapping the last 4 occurring events per attribute as regular features. Missing values were imputed for patients with less than four occurrences for the monitored attributes.

Following the principles proposed in *Chapter II-1* the assessment of classifiers follows a 10-fold cross validation, and differences in performance were tested using a t-test over the collected accuracy estimates (preserving folds) with 9 degrees of freedom. Note that the provided results in this section slightly differ from previously published results in [302] since P2MID and MaCID assume additional behavioral variations described in this section.

Results. Figure 4.4 illustrates the observed accuracy levels for the different medical tasks. The horizontal line corresponds to the default accuracy from a random learner based on the number of classes per task, $\frac{1}{|C|}$. P2MID and MaCID perform better than traditional classifiers across the selected prediction tasks. This improvement is statistical significant (at α=1%) and motivates the importance of modeling temporal and cross-attribute dependencies. P2MID shows a distinctively better accuracy against MaCID for the surgery prediction task (at α=5%). This is potentially related with the fact that the anticipation of surgeries is well modeled by specific compact sets of health-records with heightened discriminative power. Contrasting, MaCID shows higher accuracy for the anticipation of hospitalizations (at α=1%) due to the relevance of using a more broad view of clinical activity and MaCID's ability to better model stochastic uncertainty.

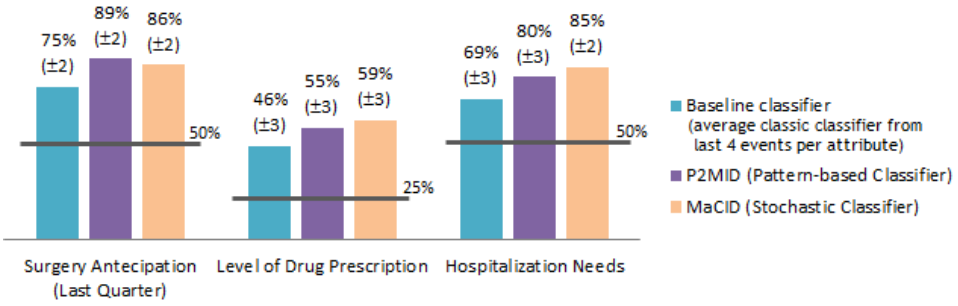


Figure 4.4: Classification accuracy of generative, pattern-based and classic learning models for three different medical tasks using a month granularity.

The efficiency of the proposed classifiers was assessed for a varying number of patients and events. Figure 4.5 gathers the results from this analysis. The time efficiency of P2MID and MaCID naturally deteriorates more rapidly than traditional classifiers. This is explained by two factors. Traditional classifiers learn from a very small subset of the overall events. Contrasting, P2MID needs to perform exhaustive searches under low levels of support and MaCID performance is penalized by the length and complexity of the underlying architectures. In terms of memory,

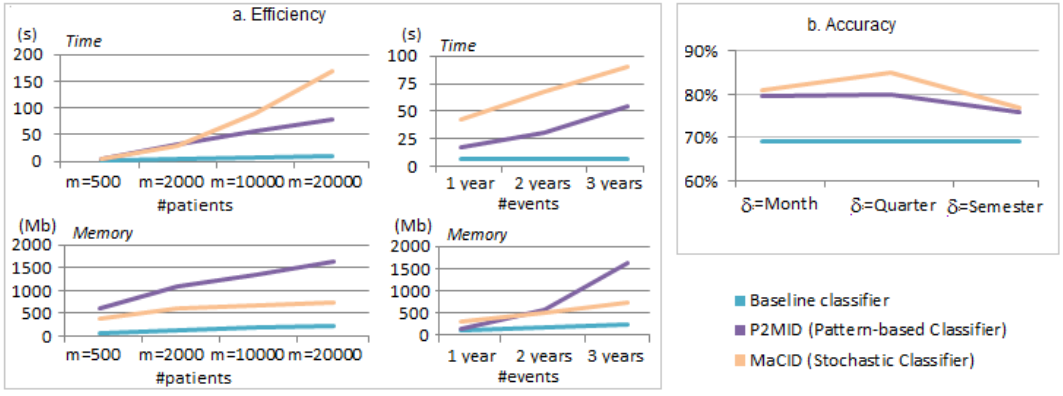


Figure 4.5: Understand the behavior of the different predictive models: *a) accuracy* for varying time scales, and *b) efficiency* for varying number of patients and events.

MaCID performs significantly better than P2MID since the learned lattices are more compact than the voluminous sets of temporal patterns found by P2MID.

Figure 4.5b compares the accuracy of P2MID and CaMID for varying levels of sparsity (number of events per time partition) by varying a fixed temporal granularity ($\delta \in \{1, 3, 6\}$) for the anticipation of hospitalization needs. The use of coarse-grained time partitions (semester granularity) is associated with the deterioration of the performance of both classifiers. For pattern-based models, the decrease on the number of partitions leads to the loss of significant precedences as they are captured as co-occurrences. For stochastic models, the behavior becomes less centered on sequential data analysis, which contradicts the original purpose of Markov-based models.

Figure 4.6 evaluates the impact of adopting time-enriched sequential patterns versus simple sequential patterns. The difference in performance is statistically significant (at α=1%). First, the proposed discriminative models tend to score preferentially patterns occurring near the time period under prediction. Also, the allowance of temporal shifts under a penalization factor during the testing stage offers a time-dependent informative context for classification. Contrasting, simple sequential patterns cannot offer temporal guarantees, and, therefore, the influence of both recent and old events to discriminate the class under prediction is not clearly differentiated. Second, the time partitioning strategy allows to deal with heightened levels of sparsity by choosing an adequate granularity with impact on the degree of precedences vs. co-occurrences.

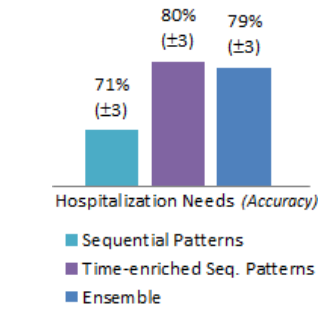


Figure 4.6: Impact of temporally enriching sequential patterns.

Finally, in Figure 4.7, the impact of using alternative HMM architectures is evaluated. The use of fully-interconnected and left-to-right architectures (LRAs) have a worse performance against SPA and MIA architectures (see Defs.IV-3.4 and IV-3.6). Since fully-interconnected and LRAs architectures have no dedicated states to emit delimiters, there is not an explicitly way of temporally aligning sequences of event-sets. Furthermore, the inclusion of delimiters is associated with convergence problems on the emission probabilities. As such, the proposed extensions create more adequate generative setting sensitive to the underlying sparsity (given by the weight of transition probabilities towards deletion or insertion states) and attribute interdependencies (given by the most probable emissions along the main path). Finally, the use of multi-path architectures is able to minimize the convergence problems associated with the event emissions when a medical condition is discriminated by multiple arrangements of events.

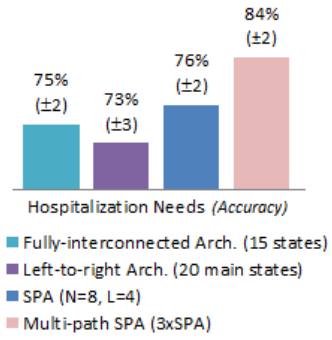


Figure 4.7: Impact of architectural decisions in MaCID.

4.4.1 Discussion

The proposed deterministic and stochastic classifiers show significant performance improvements for the adequate learning from structured data contexts. Yet, their behavior show contrasting properties that should be known before selecting the learner.

Pattern-based methods are particularly relevant when a particular class is well discriminated by small (yet highly consensual and thus statistically significant) regions. Illustrating, when predicting the need for a specific surgery, specific arrangements based on prior diagnoses and evaluations are typically sufficient to adequately discriminate this condition. Therefore, P2MID is the choice for data contexts with a large number of either non-discriminative or uninformative regions.

The proposed stochastic methods offer a more smoothed behavior since they consider all the observed elements to shape the transition-emission probabilities and rely on a probabilistic view that better models mismatches across observations due to unexpected noise. As such, they are the natural choice for more complex classification

tasks, such as the prediction of the needs for an hospitalization, which can be explained/discriminated by a high multiplicity of interrelated factors (often associated with large number of lengthy arrangements).

Accordingly, the decision on when and how to use stochastic methods depend on three major factors: 1) on the class-conditional regularities underlying data (MaCID is only required in the scarce presence of discriminative cascades or arrangements of events); and 2) on whether the relevance of the occurrences along the timeline is uneven (P2MID is able to differentiate the relevance of an event based on its time of occurrence to, for instance, prefer more recent events).

4.5 Summary of Contributions and Implications

This chapter addressed the challenging task of learning classifiers from informative and discriminative regions from (high-dimensional) structured data contexts. For this aim, the major requirements associated with this task were motivated (including the need to model temporal and integrative cross-attribute views) and the major contributions and limitations from related work towards this end were surveyed. In this context, we overviewed recent attempts towards the distance-based, pattern-based and stochastic classification of temporal data and, due to their limited ability to deal with attribute multiplicity, integrative learning principles. In order to address the observed limitations, we built upon previous contributions from *Book IV* and proposed two distinct classifiers.

First, we propose a pattern-based classifier, P2MID, for an associative learning based on discriminative, integrative and temporal patterns. For this aim, P2MID uses discriminative criteria to affect the discovery of relevant regions and defines a new integrative score of their relevance based on revised notions of support and lift. These notions are, in this context, adapted to allow for the presence of structural and temporal misalignments on the observations supporting a given temporal pattern. The allowed misalingments are also tolerated when matching a testing observation against the learned regions. P2MID was further extended to adequately handle the specificities of cascades and arrangements of events.

P2MID allows for variations on its default behavior. First, it can be easily parameterized to prioritize occurrences on certain time periods to, for instance, attenuate the discriminative impact of older events. Finally, the observed temporal mismatches between a testing observation and the learned regions can be used to proportionally affect the score.

Second, we propose a stochastic classifier, MaCID, with dedicated architectures able to focus the learning on specific regions of interest from time-enriched itemset sequences. Contrasting with P2MID, the discriminative properties of the learned class-conditional models are only assessed during the testing stage based on their likelihood to describe a new observation. This turns MaCID particularly relevant for data contexts where a class is not easily discriminated by a low number of compact regions. The customization of MaCID to support incremental learning, the modeling of numeric data, a correct interpretation of multi-item assignments, the adequate initialization of emission probabilities, and the efficient annotation of time guarantees are discussed.

In order to guarantee the applicability of these classifiers towards a large multiplicity of real-world data we guarantee an adequate mapping of multi-dimensional databases, relational databases, multi-sets of events, three-way time series into an integrative and temporal data structure. Furthermore, we showed how labels can be easily retrieved for supervised tasks and be used to shape the proposed mappings, and proposed new principles to handle data structures characterized by a mixture of temporal and static attributes.

The conducted experiments hold evidence for the accuracy and utility of the proposed associative classifiers. For this aim, we selected distinct healthcare tasks, and confronted the properties of deterministic and stochastic learners. Their performance is essentially dependent on the observed class-conditional regularities: pattern-based models are preferred in the presence of discriminative regions, while probabilistic models are preferred when a given class is better discriminated by a high number of interrelated factors or by global regularities.

5

Classification from Significant Regions

Guaranteeing the statistically significance of classification decisions is of increasing importance to validate biological and clinical markers and to support computer-aided decisions associated with medical, trading, marketing, administrative and other initiatives with either high impact on daily lives or high costs. Despite the contributions proposed along *Chapters VI-1-VI-4* to learn accurate classifiers, they are insufficient to guarantee the statistical significance of their decisions. As such, this chapter shifts the focus from the optimization of the average performance of classifiers towards their ability to adequately generalize and thus minimize the variability of performance. As a result, this chapter combines both accuracy and significance views to address the challenges associated with learning local classification models from tabular and structured data.

Guaranteeing the statistical significance of classifiers learned from high-dimensional data is challenged by two major problems. First, global classifiers are commonly applied on reduced data spaces where the significance can no longer be properly assessed. Feature selection and other forms of dimensionality reduction are used to decrease the complexity of the learning task, remove uninformative features and reduce the propensity of classifiers to overfit the input data. However, these procedures cannot flexibly identify regions of interest and introduce new forms of bias since they are mainly driven by the compactness and discriminative power of the new data space (neglecting statistical significance). Second, local classifiers commonly infer decisions from regions associated with minimal subsets of features whose joint values discriminate a class by chance. In particular, the decisions placed by state-of-the-art associative classifiers and decision trees in high-dimensional data contexts are typically inferred from non-significant regions (false positives) due to the lack of proper statistical assessments [316]. As a result, their performance typically suffers from high variability (mainly explained by the bias component of the error) due to an heightened underfitting propensity. This problem aggravates for data with a considerably low number of observations ($n<m$).

Although *Book V* proposed principles to guarantee the discovery of statistically significant regions from high-dimensional data, these principles are insufficient to guarantee their significance in labeled data contexts. A region (either a bicluster, cascade or arrangement) can be significantly informative, yet not significantly discriminate. As an attempt to address these observations, this work introduces new principles that guarantee the statistical significance of local classification models without compromising their accuracy. These principles are proposed in three major steps. First, we propose statistical tests to assess the significance of discriminative regions from labeled tabular data. Second, we incorporate these principles within the proposed learning methods. Third, we discuss their relevance not only in the context of associative classification, but also to guide the learning of other local classifiers. Finally, we extend these contributions towards labeled structured data. Accordingly, four unique contributions are made available:

- statistical tests to assess the significance of (discriminative) regions from labeled tabular and structured data;
- new class of decision trees and random forests with reduced propensity to underfit high-dimensional data;
- extended associative classifiers with bounds on the provided guarantees of significance;
- annotation of decision rules with statistical indicators of their significance to support real-world decisions.

The relevance of the proposed contributions is corroborated by initial empirical results on both synthetic and

real-valued data. Clinical data is selected for this end due to the criticality of framing the significance of computer-aided medical decisions, as well as biological and social data contexts with a limited number of observations.

Figure 5.1 illustrates the identified challenges and proposed contributions to guarantee the significance of associative decisions. The need to measure the impact of training and testing functions on the significance of decisions is tackled in the next chapter.

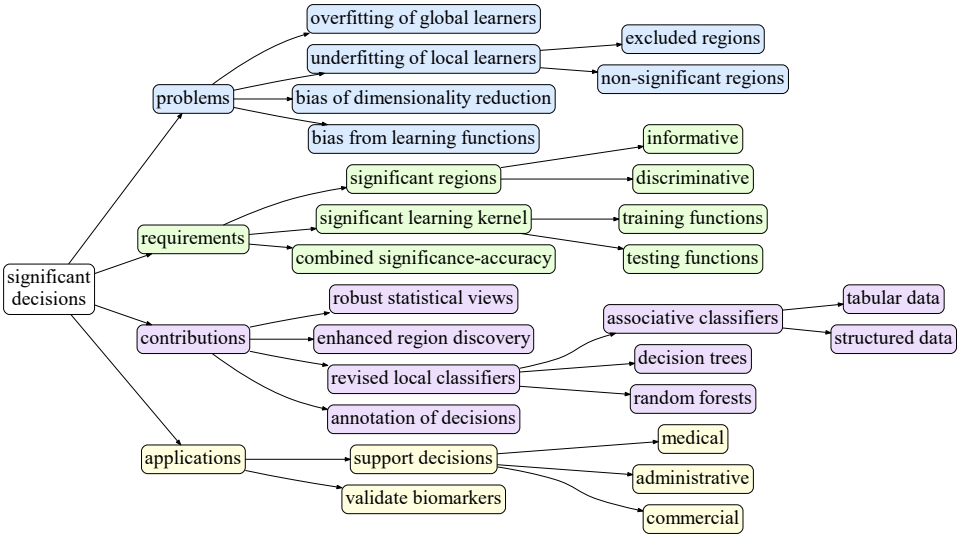


Figure 5.1: Guaranteeing the significance and relevance of associative decisions in high-dimensional spaces.

This chapter structured as follows. *Section 5.1* provides further background on the target task. *Section 5.2* describes state-of-the-art contributions. *Section 5.3* introduces the solution space. *Section 5.4* gathers experimental results that provide initial empirical evidence for the utility of these contributions. Finally, concluding remarks and implications from this work are synthesized.

5.1 Background

The impact of learning classifiers from regions of high-dimensional data needs to be properly addressed to primarily minimize their risk of underfitting data (associated with the selection of non-significant regions), but also their risk of overfitting data (associated with the selection of regions with loose homogeneity). However, the statistical significance of these regions (from which decision rules are inferred) is often disregarded. Instead, regions are selected as a consequence of an improved discriminatory power. However, in high-dimensional data spaces, it is highly probable that a small subset of the original features is able to be informative and discriminate by chance. To our knowledge, there are not yet studies able to adequately address and quantify the impact of this problem. As such, this chapter aims to experimentally measure how the significance of the modeled regions affect the performance of classifiers. For this end, we further motivate this problem and explain why the contributions proposed in the context of *Book V* are insufficient to answer it.

Challenges. The limitations associated with existing classifiers can be better understand when presented separately according to the locality of their learning functions. Global classifiers have high propensity to overfit the input data due to their inability to flexibly exclude uninformative regions [644, 200]. To minimize this problem, global classifiers can be parameterized with procedures for dimensionality reduction to model (n, p)-spaces where $p \ll m$. However, existing procedures to reduce dimensionality (including feature selection and hyper-dimensional mappings) are not sufficiently flexible to guarantee the accommodation of all relevant regions and exclusion of non-relevant regions (see *Section I-1.2*). As the majority of existing procedures reduce dimensionality by factors [682, 316], the learned global classifiers show a contrasting high propensity to underfit the original data space.

Local classifiers typically underfit the input data due to two major reasons. First, decisions are inferred from a single or few regions, such as decisions inferred from decision trees [383, 542] and random forests [96]. In this context, decisions are made in the absence of complete information due to a high propensity to exclude regions of interest [316]. Second, decisions are commonly inferred from either non-significantly informative or discriminative regions. Illustrating, consider a decision tree built over a real-valued (n=100,m=10000)-space with two balanced classes and a branch with tests on p=3 features. There is a high probability that a decision rule based on this region is discriminative by chance and thus non-significant. This problem is aggravated in high-dimensional data where the number of observations does not largely exceed the number of features (Figure I-1.18 illustrates this undesirable effect).

This chapter targets local classifiers. Yet, the proposed principles can be easily extend for the assessment of global classifiers by seeing the input data as a single (discriminative) region and by studying how their performance changes when this region shows varying guarantees of statistical significance.

Requirements. The contributions proposed in the context of *Book V* to assess the statistical significance of regions are required towards this end. However, their sole application does not guarantee the satisfaction of the introduced problems. First, not only the size of a region (given its homogeneity criterion) needs to be significant, but also its discriminative power. In other words, the significance of both the support and confidence of a region (Def.VI-1.6) must be simultaneously satisfied whenever possible to guarantee that a region is significantly informative and discriminative.

Second, not only these statistical assessments are required, but the learning functions for tabular and structured data contexts adequately revised to guarantee that decisions are inferred from a complete set of significant regions.

Finally, bounding the confidence of decisions is additionally required for labeled data contexts without significantly informative and discriminative regions for one or more classes. Note that this indicator of confidence differs from the previously introduced concept of class strength. Illustrating, a classifier with probabilistic outputs may suggest that one class has over 90% of probability to be associated with an observation, yet this decision may have low confidence if its inferred from non-significant regions.

5.2 Related Work

Despite the large extent of research on how to learn classifiers from high-dimensional data [345, 329, 105, 112, 552], they are not able to address the target problem due to four major limitations. First, the learning of classifiers is (mistakenly) seen as independent from the applied forms of dimensionality reduction [589, 343]. Understandably, dimensionality reduction can exclude relevant regions and include non-relevant regions, introducing new sources of biases to the learning task.

Second, although a high number of classifiers proposed in literature accommodate principles to guarantee a reduced propensity towards over/underfitting [173, 330, 405], these principles are in general insufficient when learning from high-dimensional data with a limited number of observations [316, 345]. Illustrating, decision trees make available pruning principles to guarantee that the learned trees do not overfit data [173]. However, their application over high-dimensional data is typically associated with a heightened underfitting propensity since the search for discriminative regions is finished for a given class as soon as a combination of features' values is able to discriminate it. To minimize this problem, random forests can be considered [405]. However, a random forest simply adds a voting stage from a larger number of (uncertain) decisions, thus not being able to adequately minimize the underfitting risk[1].

Third, the previous limitations are aggravated for classifiers learned from structured data. In particular, guaranteeing the significance of the underlying regions is a challenging problem for these data contexts (see the surveyed

[1] The use of out-of-bag error as an estimate of the generalization error is primarily used to control its propensity to overfit data.

work from *Section VI-4.2*). Furthermore, the use of classifiers based on stochastic models of data have no objective criteria to guarantee an adequate generalization [309].

Fourth, although some classifiers place statistical tests to guarantee the inference decisions rules from regions with high discriminative power, these tests are insufficient to guarantee the significance of these rules. Statistical tests of the discriminative power of a region include: χ^2 tests [401], Fisher tests [99], low inter-class overlapping tests [497], among others [99, 130]. As such, to our knowledge, there is not yet an integrated view that simultaneously tests the informative and discriminative significance of classification rules.

5.3 Solution

The target problem is answered in four major steps. *Section 5.3.1* extends the statistical tests from *Book V* to statistically assess the discriminative power of regions from labeled tabular data. *Section 5.3.2* shows how these statistical views can be used to guarantee the discovery of significant regions. *Section 5.3.3* relies on these contributions to revise the behavior of both associative classifiers and random forests, as well as to annotate decision rules with new indicators of confidence. Finally, Section 5.3.4 extends these contributions towards structured data.

5.3.1 Significance of Discriminative Biclusters: Local Assessments

This section proposes an integrated statistical assessment of the informative and discriminative power of biclusters.

Informative Regions. For the purpose of guaranteeing the statistical significance the selected regions, we rely on the statistical tests proposed throughout *Book V*. Based on these tests, two major strategies were proposed to guarantee the significance of regions (biclusters) from local descriptive models (biclustering models). First, local statistical tests can be applied to prune the solution space or post-filter non-significant biclusters. These tests rely on the computation of tails of binomial distributions (affected to the observed coherency assumption, coherency strength and quality) to calculate a (corrected) p-value indicating the probability that the support of a bicluster pattern deviates from the expectations[2]. Second, global tests can be used to infer constraints of minimum size that guarantee the significance of a bicluster. These global tests rely on the assumption that significance is verified when the number of found biclusters satisfying a minimum support (number of rows) and pattern length (number of columns) deviates from the expected number of biclusters satisfying same thresholds. Assuming that the expected number of biclusters follows a Poisson distribution (approximated from statistics on permuted data), this can be easily done by studying the thresholds where this hypothesis is rejected. Similarly to the first strategy, these global constraints can be used to filter non-significant biclusters or, more interestingly, be inputted to pattern-based biclustering to guarantee the retrieval of a reasonable number of dissimilar biclusters able to satisfy them. *Section V-1.3.3* revised pattern-based biclustering towards this end. Both of these two strategies have pros and cons for the goal of guaranteeing that the statistical significance of descriptive regions. Local statistical tests can be computationally expensive in the presence of a high number of biclusters. Global statistical tests tackle this problem and can be used to seize additional efficiency gains when the associated constraints are inputted as an heuristic to guide biclustering. Nevertheless, global tests are less precise and have a higher propensity towards false positives and false negatives for datasets with a non-uniform distribution of numeric/symbolic values.

By using one of the two previous strategies, we guarantee that a bicluster is informative, meaning that the satisfied criteria of homogeneity (coherency strength, assumption and quality) is statistically significant for the observed number of rows and columns given the global regularities underlying the input data. In the context of labeled data, this assessment can be made for a given class-conditional partition or for a set of partitions whenever there are labels in the consequent. We can thus say that a bicluster is informative with regards to a set of labels, $C \subset \mathcal{C}$, if its occurrence deviates from expectations on observations with a label in C.

[2] Statistical tests in: http://web.ist.utl.pt/rmch/software/bsig/

Discriminative Regions. A bicluster can be informative yet non-discriminative, and therefore of few relevance for an associative model. To address this problem, we also need to statistically test its relevance from a second angle: whether it is able to significantly discriminate a subset of classes.

The simplest way to guarantee that a bicluster is both informative and discriminative is to verify if its occurrence is significant for a particular subset of classes C and non-significant for the remaining classes $\mathcal{C} \setminus C$ (where $C \neq \mathcal{C}$). *Basics 5.1* provides an illustrative assessment of the discriminative significance of an observed bicluster. In FleBiC, this assessment is applied by verifying that a rule $\mathbf{B} \Rightarrow C$ has a p-value lower than 0.01 for observations with labels in C, and a p-value higher than 0.05 for the remaining observations. Illustrating, given a dataset with two balanced classes, a rule $\mathbf{B} \Rightarrow C$ from a bicluster with $\varphi_{\mathbf{B}}$ pattern may have to be supported by more than 1/3 of C observations and less than 1/5 of $\mathcal{C} \setminus C$ observations in order to be considered a true positive discovery.

Basics 5.1 Illustrative assessment of discriminative regions from tabular data

Consider a tabular dataset with m=200 features, 3 classes $\{c_1,c_2,c_3\}$ with n_1=n_2=n_3 s=40 observations, and a class-conditional discretization of features using 3 items uniformly distributed. Assume that we observed a trio of features with constant items across 10 rows for c_1, 8 rows for c_2, and 2 rows for c_3. Is this region $\mathbf{B}$ statistically significant? A simple binomial calculation shows that the probability of its support to include *at least* 18 $\{c_1,c_2\}$-observations is $p'_{\mathbf{B}|\{c_1,c_2\}}$=8.4E-11 and *at most* 2 c_3-observations is $p'_{\mathbf{B}|c_3}\approx$1. Although this probability is non-significant for c_3 and it appears to be considerably low for $\{c_1,c_2\}$, we need to consider the effect of: 1) the dimensionality (worst-case $p_{\mathbf{B}|\{c_1,c_2\}}=\binom{500}{3}p'_{B|\{c_1,c_2\}}$=1.1E-4), and 2) its deviation from expectations. Assuming the Bonferroni correction with α=0.05 significance, then $p_{\mathbf{B}|\{c_1,c_2\}}$ needs to be assessed against 8.3E-3. Thus, under these assumptions, the rule $\mathbf{B}\Rightarrow\{c_1,c_2\}$ would be considered a true positive discovery (informative and discriminative).

Although this is a fundamental criterion to guarantee the significance of a region, additional criteria can be defined to test meaningful ratios that accommodate its discriminative significance, including:

- χ^2 and Fisher tests [401, 578];
- (*normal* setting) weighted product of $p_{\mathbf{B}|C}$ with the probability of the $\varphi_{\mathbf{B}}$ to be verified *at most* n_j times for the remaining classes $c_j \in \mathcal{C}\setminus C$:

$$p_{\mathbf{B}|C} \prod_{c_j \in \mathcal{C}\setminus C} \sum_{x=0}^{q} \binom{n_j}{x} p_{\varphi_{\mathbf{B}}|c_j}{}^{x} (1 - p_{\varphi_{\mathbf{B}}|c_j})^{n_j - x} < \alpha \tag{5.1}$$

- (*relaxed* setting) ratio of the weighted conditional probabilities of occurrence:

$$p_{\mathbf{B}|c} \prod_{c_j \in \mathcal{C}\setminus C} \frac{1}{p_{\mathbf{B}|c_j}} < \alpha \tag{5.2}$$

Basics 5.2 provides an instantiation of these metrics for an illustrative dataset.

Basics 5.2 Statistical views on the discriminative power of an illustrative bicluster

Consider the discrete tabular dataset provided in Figure 5.2 with $|\mathcal{L}|$=5, $|\mathcal{C}|$=3 and the colored non-perfect constant bicluster $\mathbf{B}$. According to the principles proposed in this section, $p_{\mathbf{B}|c_1}$=2.0E-5, $p_{\mathbf{B}|c_2}$=2.0E-5 and $p_{\mathbf{B}|c_3}$=7.4E-2. Assuming a Bonferroni correction, these levels need to be assessed against 4.2E-4 at α=0.01 (significance max-threshold) and 2.1E-3 at α=0.05 (non-significance min-threshold). As such, $\mathbf{B}\Rightarrow\{c_2,c_3\}$ is significant. To compute a single ratio incorporating the discriminative criteria $p_{\mathbf{B}\Rightarrow\{c_2,c_3\}}\approx$4.8E-7 according to (5.1) and $p_{\mathbf{B}\Rightarrow\{c_2,c_3\}}\approx$6.4E-6 according to (5.2).

	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$	$\mathbf{y}_5$	$\mathbf{y}_6$	...	$\mathbf{y}_{20}$	class
$\mathbf{x}_1$	1	2	4	4	6	2	...	4	c_1
$\mathbf{x}_2$	5	1	2	5	2	3	...	3	c_1
$\mathbf{x}_3$	3	2	**3**	**2**	**4**	**5**	...	3	c_1
$\mathbf{x}_4$	2	4	1	5	1	4	...	2	c_1
$\mathbf{x}_5$	2	3	1	4	6	7	...	3	c_2
$\mathbf{x}_6$	4	3	**3**	**3**	**4**	**5**	...	5	c_2
$\mathbf{x}_7$	5	2	**3**	**3**	**3**	**5**	...	4	c_2
$\mathbf{x}_8$	1	4	**3**	**3**	**4**	**5**	...	3	c_3
$\mathbf{x}_9$	3	5	**4**	**3**	**4**	**5**	...	2	c_3
$\mathbf{x}_{10}$	2	1	1	2	1	2	...	3	c_3

Figure 5.2: Discriminative region from a labeled dataset.

5.3.2 Using Significance to Shape Associative Models

Under the enlarged notion of statistical significance, new principles are required to guarantee an adequate discovery of significant regions from labeled (tabular) data. In line with the contributions discussed in *Section VI-5.3.1*, two major strategies are proposed for this end. First, biclusters can be discovered for each class-conditional under a

minimum support that is able to guarantee a low number of false negatives, i.e., if a bicluster is not discovered there is a high probability of being not significant. Based on the recovered biclusters, the principles proposed in *Section VI-1.3.1.3* for the efficient composition of rules with disjunctions of labels in the consequent are extended with local statistical tests to verify whether a bicluster is significantly informative and discriminative (by statistically testing its support for each class).

Second, minimum pattern support and length constraints can be inferred from global statistically testing each class-conditional data partition, resulting in a total of $|C| \times 2$ constraints. These constraints are then used to pruned the solution search space size, and the search is iterated until a minimum number of biclusters (or data space coverage) is found while guaranteeing the satisfaction of these constraints.

Scarcity of Simultaneously Informative and Discriminative Regions. A given dataset may only show a compact set of statistically significant biclusters due to the scarcity of regions with homogeneity, size and discriminative power deviating from expectations. To minimize this problem, we propose the use of the learning functions introduced in *Chapter III-9* (and extended in *Section VI-1.3.1*) due to their ability to assume a flexible positioning, varying coherency and quality (tolerance to noise), and thus be able to detect less trivial yet significant biclusters.

Nevertheless, when the input data space does not contain regions with flexible homogeneity satisfying the significance criterion, one of the two following principles can be used. First, the local and global tests should rely on more relaxed statistical thresholds (never less than α=0.1 though).

Second, and whenever the relaxation of the statistical power is not sufficient to retrieve a considerable number of informative and discriminative regions per class according to the input stopping criteria, we suggest the output of the top K biclusters according to their statistical p-values. By default, we consider the output of the top K=5 decision rules per class c, including decision rules with multiple classes in the consequent C (where $c \in C$). For the few datasets where the application of this principle may be necessary, the non-significant rules should be clearly annotated with the tested significance, and a disclaimer should be considered on classification decisions from observations matching some of the regions of these rules.

5.3.3 Enhancing Local Classifiers for Tabular Data

Let us now discuss the implications that the proposed principles have in the behavior of local classifiers. For this aim, *Section 5.3.3.1* extends the previously proposed associative classifiers in this book, while *Section 5.3.3.2* revises decision trees and random forests. Nevertheless, the proposed principles can be alternatively used to enhance other classifiers, including logistic model trees[3] [383], PART rules[4] [229], decision tables[5] [373], among others.

5.3.3.1 Associative Classification

The learning schema for *associative classification* proposed in *Chapter VI-2* is here enhanced with the principles from previous section. In this way, the discovery of regions with flexible coherency and quality further provides guarantees of statistical significance. The lines 11, 13 and 20 of Algorithm VI-10 are revised to guarantee the accommodation of the previous tests, which may impact the total number of iterations in order to guarantee that a reasonable number of rules per class is discovered.

We further propose a variant of the integrative scoring of regions defined by Eq.(VI-2.1) to accommodate the statistical significance criteria. In this context, a new component is included with weight $0.3 \times \text{sig}_R + 0.7 \times \omega_R$ where sig_R measures the discriminative significance of rule R and according to the score table $\{1{:}p \leq 0.002,\ 0.7{:}0.002 > p \geq 0.01,\ 0.4{:}0.01 > p \geq 0.1,\ 0{:}p > 0.1\}$, where p is the statiscal p-value according to Eq.(5.1). Finally, the testing function is preserved. This enhanced classifier is referred as FleSBiC (Flexible and Significant Bicluster-based Classification).

[3] Classification trees with logistic regression functions at the leaves.
[4] Classifier that builds a partial C4.5 decision tree in each iteration and makes the best leaf into a rule.
[5] Tests on subsets of features better discriminating each class: one region per class spanning the space of all observations.

5.3.3.2 Decision Trees and Random Forests

Enhanced Decision Trees. A decision tree defines a tree structure where the nodes correspond to tests on the features from a region of interest, the branches represent the outcome of a decision, and leafs define the outcome (class). Each path from root to leaf represents a decision rule and thus defines a discriminative region (bicluster). Figure 5.3 illustrates how a labeled tabular data is partitioned into smaller regions to compose the target tree.

Decision trees are chosen since they rely on regions with a low number of features. The search for additional features stops whenever the combination of values for a given set of features is already able to separate two or more classes, independently of whether they are discriminative by chance or not. In these contexts, decision trees largely underfit the input data (decisions commonly made from less than 5% of features for m>200) and therefore they show a high variability of error between cross-validation folds.

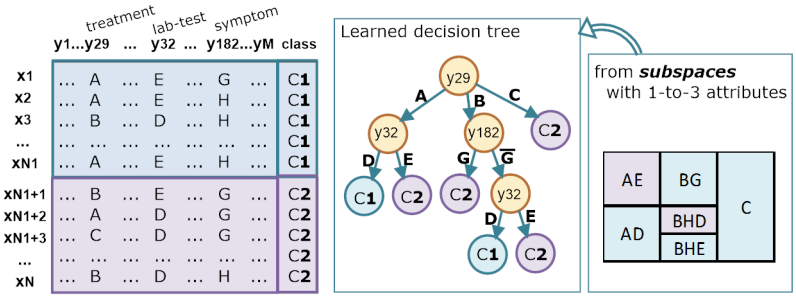


Figure 5.3: Decision trees are inferred from regions of interest and thus emulate the behavior of associative classifiers.

In order to assess the significance of each decision rule, $\mathbf{B}_k \Rightarrow C$, the associated regions need to be recovered from the learned tree and either individually tested or verified against constraints inferred from global tests (using the principles introduced in *Section 5.3.1*). Leafs can be annotated with a measure of significance to offer a barometer of the confidence of the decision. In this fashion, the revised decision trees output a pair (class,confidence).

The proposed assessment triggers new implications for the extension of local classifiers: non-significant c_i-conditional regions should be either removed or replaced by new significant rules as long as there is proof for their existence. When considering decision trees this can be implemented using the two following principles.

First, non-significant classification rules (that is, tree paths with false-annotations on leafs) should be extended to include more attributes as long as they preserve discriminative significance. This can be done by iteratively adding attributes with the highest information gains until the grown region becomes significant.

Second, when an original or grown decision rule is not significantly discriminative (often when a region becomes significantly informative for more than one class), the decision rule should be removed in an attempt to identify a new one. To implement this principle, the path must be signaled, the root-feature on that path removed from the set of candidate features, and a new path must be built using the standard behavior of the classifier with the reduced set of candidate features. A link between the non-significant leaf node and a new root node becomes active. In the worst-case, attributes are iteratively removed until no significantly discriminative path can be found. In this case, the original path can be maintained and its leaf is properly annotated as non-significant, meaning that there are no significant regions for the the class associated with this rule. The behavior of this extended decision tree is illustrated in Figure 5.4.

Enhanced Random Forests. Despite the benefits of the proposed behavioral changes to minimize the underfitting propensity of decision trees, they still suffer from an additional problem: decisions are inferred from a single region. In order to address this problem, thus avoiding missing additional relevant subsets of features with discriminative power, we propose suggest the generation of multiple (randomly seeded) decision trees according to the learning principles of random forests.

A random forest [96] is an ensemble learning method for classification that constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes of the individual trees. Similarly

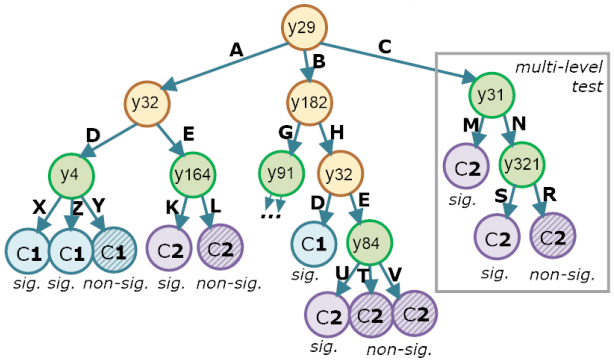


Figure 5.4: Extended decision trees where nodes are iteratively added to promote the significance of decisions.

to decision trees, a random forest can accommodate the proposed principles to provide guarantees of statistical significance on the underlying regions associated with the decision trees in the forest. Furthermore, and contrasting to decision trees, it infers decisions from multiple regions (one per decision tree in the forest), thus reducing the risk of underfitting. In this context, random forests do not only reduce the propensity of decision trees to overfit the training data (their original purpose [96]) but also the underfitting risk.

Despite the efficiency and inherent simplicity of local classifiers based on decision trees, they suffer from a lack of flexibility with regards to the composition of regions from discriminative features. First, discriminative power is assessed for all the observations or observations conditioned to a specific set of values (see entropy criteria in *Basics I-1.11*). Second, the use of information gain ratios based on entropy leads to constant regions. As such, many non-trivial yet meaningful and relevant regions (varying coherency assumption and strength) are excluded. This problem is further aggravated when the number of constant regions with statistical significance is scarce.

5.3.3.3 Annotating Classification Models

Regions from associative models and leafs from decision trees can be annotated with a measure of significance to offer a barometer of the confidence of the inferred decisions. This measure can be either given by: 1) the observed significance for C-conditional data partition $p_{\mathbf{B}|C}$, 2) by the ratios given by Eq.(5.1) or Eq.(5.2), 3) by a vector of values $\{p_{\mathbf{B}|c_i} \mid c_i \in C\}$, or 4) with a Boolean flag indicating whether it is significant for C and non-significant for $C \backslash C$. In this context, in the presence of a testing observation, decisions made by associative classifiers or random forests are framed by the significance of the underlying matched regions. This creates an informative environment to better support medical, financial, administrative and commercial decisions.

5.3.4 Enhancing Local Classifiers for Structured Data

The classifiers proposed in *Chapter VI-4* can be similarly revised to guarantee that the decisions are inferred from significantly informative and discriminative regions. In this chapter, deterministic and stochastic classifiers were proposed to learn from three-way time series and multi-sets of events. When considering the deterministic classifiers, three major principles should be considered to foster the significance of their decisions.

First, the principles to extend the statistical assessment proposed in *Section 5.3.1* can be similarly applied to assess the significance of the target regions, yet parameterized with the dedicated statistical tests proposed in *Chapter V-4*. In this context, the discriminative significance of a given cascade or arrangement of events can be easily assessed by verifying if its occurrence is significant for a subset of classes and not significant for the remaining classes or by applying the listed scores. Only the $p_{\mathbf{B}|C}$ calculus differs.

Second, the class-conditional searches for temporal patterns (from which regions given by cascades or arrangements are derived) can be preserved. However, the composed decision rules from these patterns need to be tested according to the extended statistical assessment synthesized in the previous paragraph. Similarly to the principles introduced in *Section 5.3.2*, in the absence of regions with distinctive informative and discriminative power, the

cut-off thresholds of the proposed statistical tests can be relaxed or a parameterized number of temporal patterns per class with the highest significance outputted (and their impact on the confidence of decision rules quantified).

Finally, and similarly to the revised scoring schema proposed for associative classifiers from tabular data, the integrative scoring schema for the discovered regions (given by the weighted support, length and lift of temporal patterns) proposed in *Section VI-5.3.3.1* can be weighted by the significance of the composed rules from these regions. This change is relevant since statistical significance can be loosely correlated with the weighted support and lift of a rule (Def.VI-4.2).

When considering the proposed stochastic learning variants for classification (*Section VI-4.3.3*), an important question is whether or not can we enhance them to promote significant decisions. When the proposed Markov-based models are simply used to model and decode regions from the learned models (being associative classifiers learned from them), the three previous principles are applicable. Alternatively, when decisions are directly inferred from class-conditional Markov-based models to test the likelihood of a new observation, new considerations need to be made. On one hand, the proposed Markov-based classifiers have lower propensity to underfit data in comparison with the peer associative classifiers as they explore less-probable paths in the learned lattices (associated with less relevant regions in data) to support the final decisions. As such, their decisions are inferred from a broader set of data elements from the input data space. However, on the other hand, there is no way of objectively assessing the significance of their decisions. Furthermore, although the number of iterations and convergence criteria can be modify in order to balance the propensity of Markov-based classifiers to either overfit or underfit the input data, there is no clear of adapting their behavior to provide guarantees of statistical significance.

5.4 Results and Discussion

Results are organized as follows. First, we study the fundamental properties of significant regions from labeled data. Second, we explore the limitations of local classifiers by exposing decision rules learned from real-world data. Third, we compare the performance of the extended associative classifiers, decision trees and random forests against non-enhanced classifiers. The proposed statistical tests, FleSBiC and enhanced tree-based classifiers[6] were implemented in Java (JVM v1.6.0-24) and run with an Intel Core i5 2.80GHz with 6GB of RAM.

Evaluation Settings. We selected 4 gene expression high-dimensional datasets[7] with case-control observations for the classification of leukemia (m=7129 features and n=72 observations), distinct types of lymphoma (m=4026, n=96), embryonal tumours outcome (m=7219, n=60), and colon cancer (m=2000, n=62). The assessment of the classification models over these datasets followed the principles proposed in *Chapter II-1*.

Significance of Discriminative Regions. In order to understand the properties that turns a region simultaneously informative and discriminative for a given tabular dataset, we conducted a theoretical analysis from the application of the statistical tests proposed in this chapter. This analysis, illustrated in Table 5.1 shows how the discriminative significance of a bicluster is assured based on: its properties support across classes, length, pattern, coherency assumption and strength; and the properties of the input matrix (size, dimensionality and regularities). For this end, we assumed the presence of two balanced classes, and selected the (5.1) statistical ratio in order to understand the minimum number of c_1-conditional observations (n_1) that a bicluster needs to have to be informative for a class c_1 and, complementarily, the maximum number of c_2-conditional observations n_2 that need be verified on the remaining class c_2 in order to guarantee that it is discriminative. We further assumed $p_{\varphi_{\mathbf{B}}}$ to have items with average, above-average and below-average probability of occurrence. Illustrating, a constant bicluster with pattern length $|\mathbf{J}|$=5 and coherency strength given by $|\mathcal{L}|$=5, observed in a dataset with $|\mathbf{X}|_{c_1}|$=$|\mathbf{X}|_{c_2}|$=20000 observations and $|\mathbf{Y}|$=100 features, should satisfy the following inequality $n_1 > 20 \wedge n_2 < 10$ for below-average probability and

[6] Software available in http://web.ist.utl.pt/rmch/research/software
[7] http://eps.upo.es/bigs/datasets.html

inequality $n_1 > 20 \wedge n_2 < 10$ for an above-average probability of occurrence. The analysis of the table suggests the n_1 and n_2 observations are considerably affected by all the studied parameters.

An implication from this analysis, is that the selected (5.1) ratio provides a relaxed setting to assess discriminative significance. The bounds given by the minimum and (specially) the maximum number of rows are loose, as it can be easily verified based on the differences between n_1 and n_2. The application of the alternative (5.2) ratio is associated with a decrease on the thresholds for the maximum number of observations n_2. Note, however, that both options have pros and cons when learning associative classifiers. On one hand, when requiring very large differences on the number of supporting observations $(n_1 - n_2)$, the learning algorithm may not be able to identify a large number of regions satisfying this setting. This may lead to a scarcity of matches during the testing stage. On the other hand, by allowing less delineated differences, regions are susceptible to have a looser discriminative power, thus introducing unnecessary biases for the learning task.

$\|\mathcal{L}\|$ $\|\mathbf{X}\|$ $\|\mathbf{Y}\|$ $\|\mathbf{J}\|$	5 100 100 5	5 1000 100 5	5 20000 100 5	5 200 100 5	5 200 1000 5	5 200 20000 5	10 200 100 5	5 200 100 3	5 200 100 8
Constant $p_{\varphi_B}=\frac{1}{\|\mathcal{L}\|}^m$	n_1>7 n_2<3	n_1>12 n_2<7	n_1>35 n_2<23	n_1>8 n_2<4	n_1>11 n_2<8	n_1>14 n_2<10	n_1>5 n_2<2	n_1>17 n_2<9	n_1>5 n_2<3
Constant $p_{\varphi_B}=\frac{1.2}{\|\mathcal{L}\|}^m$	n_1>9 n_2<4	n_1>16 n_2<9	n_1>63 n_2<40	n_1>11 n_2<5	n_1>14 n_2<9	n_1>18 n_2<11	n_1>6 n_2<3	n_1>22 n_2<12	n_1>6 n_2<3
Constant $p_{\varphi_B}=\frac{0.8}{\|\mathcal{L}\|}^m$	n_1>6 n_2<3	n_1>10 n_2<5	n_1>23 n_2<15	n_1>7 n_2<4	n_1>9 n_2<6	n_1>12 n_2<8	n_1>4 n_2<2	n_1>14 n_2<7	n_1>4 n_2<2
Additive ($\gamma \in [0, 0.4]$)	n_1>9 n_2<5	n_1>16 n_2<10	n_1>62 n_2<46	n_1>10 n_2<6	n_1>14 n_2<9	n_1>18 n_2<12	n_1>5 n_2<3	n_1>40 n_2<26	n_1>5 n_2<3
Order-Preserving	n_1>15 n_2<9	n_1>38 n_2<27	n_1>267 n_2<237	n_1>18 n_2<12	n_1>22 n_2<17	n_1>30 n_2<22	n_1>18 n_2<12	n_1>71 n_2<55	n_1>6 n_2<4

Table 5.1: Impact of coherency strength and assumption (and data size and dimensionality) on the minimum and maximum number of rows per class that guarantees that a region is significantly informative and discriminative (according to Eq.(1.1)). Data is balanced and the pattern of the assessed regions has average, below average and above average probability to occur (similarly to Table V-2.2).

Addressing the Underfitting Problem. Figure VI-5.5 illustrates the decision trees learned by C4.5 [542] for each one of the four gene expression datasets (on some of randomly selected cross-validation folds). Two major observations can be retrieve. First, C4.5 shows a high underfitting propensity since it only selects between 1 and 3 genes/features from datasets with thousands of genes/features. In this context, it is clearly visible that (potentially) relevant genes are excluded from the decision process. Whenever a testing observation shows values near the binarization boundary of a given testing feature, there is a high uncertainty associated with the chosen path (or outputted decision). In this context, it is clear that testing the expression-levels/values of complementary genes/features could help to guide the learning.

Second, we can see that different genes/features are selected for different folds from a single dataset. This is indicative that the phenotypes/classes are not only discriminated by a few genes, but possibly explained by complex regulatory behavior involving multiple genes. This provides further evidence for the undesirable underfitting propensity of decisions and, subsequently, of random forests.

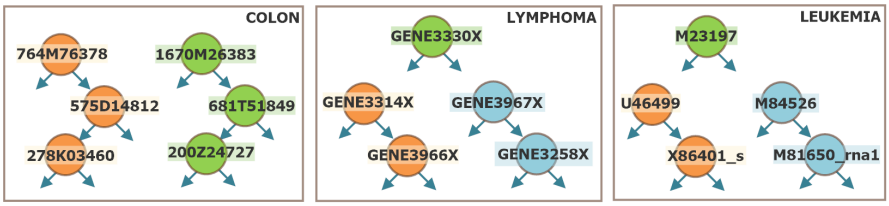


Figure 5.5: Decision trees learned for (some of the cross-validation folds) of colon, lymphoma and leukemia datasets using C4.5.

The application of the proposed principles to guarantee the statistical significance of the subset of the paths in the tree led to lengthier trees. Illustrating, for the colon dataset, the revised classifier generated decision trees with a depth ranging from 6 to 8 features (depending on the cross-validation fold). For this dataset, there were no leaves found below a depth of 4 features ($|\mathbf{J}| \geq 4$). This contrasts with the original behavior of decision trees

where some decisions are commonly inferred based on 1 or 2 features. It is clear that revised decision trees are less propense to the risk of underfitting. Note, however, that the comparison of their accuracy levels (Figure 5.7) only showed significant differences for Leukemia and Lymphoma datasets (at α=0.5). This seems to be explained by the structural noise present in the dataset. An in-depth analysis of the variables explaining these results is considered to be a priority for future work.

Impact of Statistical Significance in the Performance. To understand the impact of selecting regions with varying statistical significance on the performance of local classifiers, we adapted the behavior of FleSBiC to be able to learn biclusters with parameterizable informative significance (within an inputted range). For this analysis, we impose FleSBiC to find a minimum number of 5 discriminative biclusters per class. This guarantees that we are able to compare the performance of associative models that discover a similar number of regions, yet with varying statistical significance with regards to the regions' support. Figure 5.6 gathers the results of this analysis for the colon dataset. Figure 5.6 further shows the impact of implementing the same conditioning on the enhanced tree-based classifiers by including/removing nodes per path until the underlying regions satisfy the inputted criteria of statistical significance. Three major observations can be retrieved. First, the proposed changes in the behavior of tree-based classifiers resulted in performance improvements (reasons already explored in the previous analysis), although their average accuracy is slightly lower than the proposed associative classifier.

Second, the performance of FleSBiC's variant clearly deteriorates in the presence of biclusters without guarantees of statistical significance (>1%). This is explained by two major factors: 1) biclusters tend to have a lower number of supporting observations and thus the observed discriminative pattern is not verified on a high number of observations (including some testing observations), and 2) since the probability of a bicluster occur by chance is high, so it is the probability of being discriminative by chance.

Third, the performance slightly deteriorates when imposing regions to have high levels of statistical significance (p-values below 1E-10%). In order to satisfy these thresholds, the biclusters need to assume the presence of a large amount of noise, thus being associated with low quality regions (loose homogeneity). In this context, there is a higher probability of such biclusters being discriminative by chance. In fact, a closer analysis of the lift from the learned rules, reveals that their discriminative power is not as heightened as the rules discovered for regions with a p-value >1E-10%.

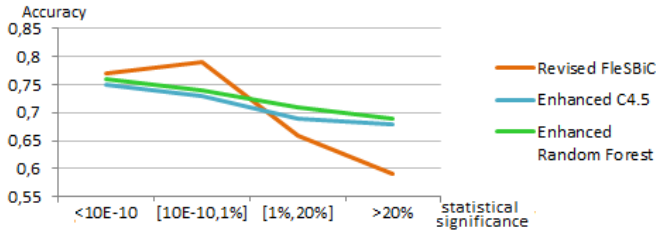


Figure 5.6: Comparison of the accuracy of three local classifiers able to learn from regions that satisfy an pre-specified range of statistical significance. The associative classifier (revised FleBiC) is prepared to learn from >5 discriminative biclusters per class.

These observations stress the importance of guaranteeing the statistical significance of regions (minimizing underfitting propensity), yet simultaneously avoiding its blind optimization (minimizing overfitting propensity).

Significance of Discriminative Regions. Figure 5.7 compares the accuracy of different local classifiers with and without guarantees of statistical significance for the colon, lymphoma, leukemia and embryo datasets. For this aim, we selected FleBiC, C4.5 and random forests in the absence and presence of the principles proposed throughout this chapter. Although it is clearly visible that the results improve consistently across classifiers learned from significant regions, not all improvements were significant (at α=0.1). In this context, there is the need to analyze the causes for the high variability of the observed error estimates across folds. The variability is slightly reduced for the enhanced classifiers, yet considerably high. For this end, the experimental analysis conducted in the next chapter

(*Section VI-6.2*) decomposes this error and relies on smoothing factors on the error in order to reveal the true performance of the classifiers before and after enhancements.

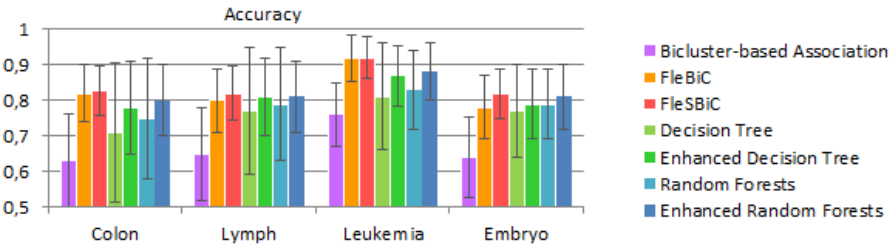


Figure 5.7: Accuracy of FleBiC, C4.5 and random forests (with and without guarantees of statistical significance on the modeled regions) for the colon, leukemia and lymphoma datasets.

5.5 Summary of Contributions and Implications

This chapter addressed the task of promoting statistically significant decisions by learning associative classification models from significantly informative and discriminative regions. The relevance of this task for high-dimensional data contexts was motivated and the major limitations of existing research discussed. To answer this task, we revised the previously proposed statistical tests to assess the significance of regions from (single-label) tabular and structured data towards multi-label data contexts. Second, we guarantee their effective incorporation to guide the discovery, composition and scoring of regions. In this context, we extended existing associative classifiers from tabular and structured data in order to guarantee that decisions are adequately inferred from significant regions. Furthermore, we generalize these principles towards alternative local classifiers. In particular, we revise the behavior of decision trees to minimize its underfitting propensity and motivate its use to adequately learn random forests and logistic model trees.

We gather initial empirical evidence that shows that the inferred classification rules across distinct data domains are commonly non-significant. Also, we further explore the properties that turn a given region statistically significant. Finally, results from real data with limited number of observations show that the enhanced classifiers are able to preserve and sometimes increase the original levels of accuracy, and have a less-variable performance. As such, the experiments support the relevance of the proposed principles to minimize the underfitting risk of local classifiers.

On a concluding note, the proposed statistical views and their use within local classifiers can be used to improve the performance of classifiers, as well as to provide a simple yet robust frame to evaluate the increasing number of implications available in literature from their application in real data, validate biological and clinical markers, and support medical, trading, administrative and commercial decisions.

Future Work. This work opens new directions for future work. First, we expect to extend the proposed experimental assessment towards different real-world data domains in order to exhaustively quantify the changes in performance from inferring decisions from statistically significant regions. In line with this direction, we expect to derive statistics from a large sample of research articles in the bioinformatics field to further support the relevance of the proposed and upcoming contributions on this matter.

Second, we aim to extend this analysis towards alternative classification models, such as support vector machines, neural networks and Bayesian classifiers.

Third, the proposed principles can also be used as a pruning heuristic to narrow the search space and leverage the efficiency of existing classifiers.

Finally, we expect to complement the proposed significance tests with additional quality criteria in order to correctly assess the relevance of classification rules derived from regions with varying tolerance to noise.

6

Learning Significant and Accurate Decisions

Learning from significantly informative and discriminative regions from high-dimensional data is essential to reduce the underfitting propensity of local classifiers, as well as of global classifiers reliant on dimensionality reduction procedures. However, learning from such regions is insufficient to guarantee statistically significant decisions since training and testing functions may introduce additional forms of bias. For instance, the use of pruning procedures during the training stage and matching relaxations during the testing stage often interfere with the guarantees of significance. As such, it is critical to measure and minimize the negative impact of the selected learning schema on the significance of classification decisions.

By addressing this requirement, the learning of classifiers becomes centered on outputting statistically significant decisions. Nevertheless, the blind optimization of significance guarantees can degrade accuracy levels. For instance, large discriminative regions with loose homogeneity can promote significance yet may be of residual value to the classification task. In this context, this chapter tackles an additional requirement: integrating accuracy (average error) and significance (variability of error) views.

This chapter tackles the two previous requirements, thus offering an integrative view of the contributions provided throughout this book. As a result, it provides three major contributions:

- structured view on the impact of training and testing functions on the guarantees of significance;
- principles to promote significant decisions;
- enhanced classifiers with behavior oriented to jointly optimize significance and accuracy;

Experimental results support the aforementioned contributions, providing an overall assessment on the distinctive properties of the proposed classifiers (against state-of-the-art classifiers) in high-dimensional data contexts. These results are essential to determine the validation of the thesis hypothesis.

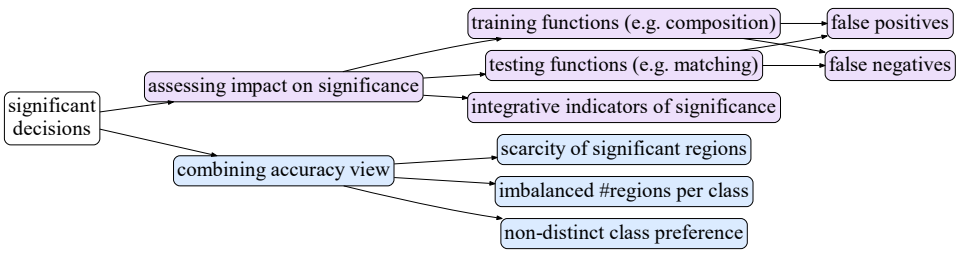


Figure 6.1: Contributions to promote significant decisions without compromising accuracy.

Figure 6.1 synthesizes the proposed contributions. According to this figure, *Section 6.1* describes the solution space. *Section 6.2* assesses the gains in performance of the enhanced classifiers against peer classifiers. Finally, we provide a summary of the contributions and implications of this work.

6.1 Solution

Below, we provide principles on how to assess and optimize local classifiers based on their learning scheme. *Sections 6.1.1* and *6.1.2* quantify the impact of training and testing functions on the guarantees of significance. *Section 6.1.3* provides meaningful scores of the significance of classification decisions. Finally, *Section 6.1.4* enhances local classifiers with principles to combine significance and accuracy views to guarantee a robust learning for non-trivial data contexts.

6.1.1 Significance of Training Functions

The training step of local classifiers can degrade the guarantees of significance of the underlying regions (thus impacting the significance of the outputted decisions) in two different ways: by either incurring in false positives or false negatives.

On one hand, the application of pruning procedures with impact on the properties of the underlying regions is often an inherent part of the training step that can increase the false positive rate (inferring decisions from non-significant regions). Illustrating, the branches of decision trees are commonly pruned to minimize their propensity to overfit data, yet this increases their underfitting risk and decreases the number of features per region thus decreasing their statistical significance (increasing risk of occurring/discriminating by chance). In this context, to measure the impact of similar procedures, the statistical significance of the new pruned regions needs to be reassessed and overwrite the previously computed p-values.

On the other hand, the application of pruning procedures with impact on the composed structure and/or number of regions can increase the false negative rate (neglecting significant regions when making decisions). Illustrating, pruning procedures applied in associative classifers, such as CMAR [401] and the proposed FleBiC, remove of certain regions in order guarantee a balanced number of discriminative regions per class. Understandably, the removed regions can be statistically significant and therefore of added value to infer more informed (and consequently less biased) decisions (see Figure VI-1.2). In this context, in order to measure the impact of removing low scored (yet significantly informative and discriminative) regions, there is the need to account for the extent (in terms of number and relevance) of lost significant regions. For this aim, there is the need to compare the selected regions for classification with regions discovered under a search that is able to provide four guarantees: 1) flexibility (flexible structures and homogeneity), 2) adequate space exploration (ideally under strict optimality guarantees), 3) dissimilarity guarantees between regions, and 4) significance of the output regions under a controlled false negative rate (non-conservative corrections). As BicPAMS satisfies these criteria, it can be use to measure the incurred loss of relevant regions. Under this knowledge, the significance guarantees of the classifiers can be linearly or squarely penalized by the degree of undesirably lost regions.

6.1.2 Significance of Testing Functions

Similarly to training functions, testing functions can also impact the significance guarantees of the outputted decisions by either increasing the false positive rate or false negative rate.

On one hand, when the matching criteria is too restricted (intolerant to noise), it can cause the loss of significant regions, thus increasing the underfitting risk by increasing the number of false negatives (missed significant regions for a testing observation). Illustrating, associative classifiers that require the exact matching between a region and a given observation often miss relevant information in the presence of noise.

On other hand, when the matching criteria is excessively relaxed (high tolerance to noise), it can be associated with spuriously matched regions, thus increasing the overfitting risk by increasing the number of false positives (considering non-relevant regions for a testing observation).

In order to guarantee an adequate measure of the previous two factors on the significance guarantees of a given classifier, there is the need to fix an adequate matching threshold, γ, to minimize both risks. According to Def.VI-1.8,

FleBiC fixes the threshold at γ=80%. Regions with matching above 80% are included for decision making (yet their score is squarely penalized according to the amount of tolerated noise), while regions with matching below 80% are discarded. In this context, the impact of either loosing significant regions or accommodating non-significant regions can be assessed against a parameterizable γ threshold for a given domain.

6.1.3 Indicative Significance of Classification Decisions

Under the introduced criteria, an indicator of the significance of classification decisions, $\phi \in [0, 1]$, can be computed. If a decision is inferred from a complete set of statistically significant regions/rules, the decision is considered to be statistically significant. If it is inferred from both significant and non-significant regions, the fraction of statistically significant regions can be used as an indicator of the decision's significance. Finally, if it is inferred from an incomplete set of significant regions, the fraction of lost regions can be either used as an alternative indicator or multiplied with the previous indicator (fraction of significant regions) for a fair assessment.

The previous indicators can be weighted by the relevance of the matched rules for a placed decision, where their relevance is given by the integrative score ω_R (according to Eq.(VI-2.1)). *Basics 6.1* provides an illustrative context for the calculus of such indicator of a decision's significance.

Basics 6.1 Computing an indicator of the significance of a decision

Consider a labeled tabular dataset with three c_1-conditional decisions rules: $R_1 : \mathbf{B}_1 \rightarrow c_1$, $R_2 : \mathbf{B}_2 \rightarrow \{c_1, c_2\}$ and $R_3 : \mathbf{B}_3 \rightarrow c_2$, with significance levels (according to (5.1)) p_{R_1}=0.1, p_{R_2}=0.002 and p_{R_3}=0.007, and with integrative scores ω_{R_1}=0.6, ω_{R_2}=0.4 and ω_{R_3}=0.3. Given an unlabeled observation whose values match with the pattern of these three regions, and a classifier that infers a decision based uniquely on $\{R_1, R_3\}$ rules. Since the significant R_2 rule is neglected and the non-significant R_1 rule is considered, the decision has an indicative significance score of $\frac{1}{3}$. For a more fair computation of the significance, we can weight by the region's relevance. In this context the significance score is $\frac{0.3}{0.6+0.4+0.3} \approx 0.23$.

6.1.4 Combining Significance and Accuracy Views

Despite the relevance of promoting significant decisions, orienting the learning uniquely towards significance criteria may have undesirable effects on the accuracy of classifiers. Illustrating, a classifier that neglects non-significant regions may end up with a small set of significantly discriminative regions per class, which may lead to a scarce number of matches during the testing step. In this context, we provide three principles to guarantee a joint optimization of significance and accuracy views.

First, in the presence of a low number of significantly discriminative and informative regions per class, the classifier should be able to accommodate additional regions (even if their significance is only satisfied under a lower α-threshold). By default, when a local classifier is not able to discover more than 5 significant regions/rules (including regions with disjoint consequents) per class, non-significant regions can be accommodate. This behavior minimizes the uncertainty associated with the lack of matched regions to infer decisions, thus increasing accuracy.

Second, in the presence of an uneven number of rules per class, instead of pruning regions (possibly degrading significance) or maintaining the imbalanced structure (possibly degrading accuracy), the candidate regions for pruning are simply flagged but not removed. As such, the flagged regions are only discarded during the testing stage whenever there are matches for two or more classes, and those classes are associated with distinct number of flagged regions (by default, when the difference between the class with the least and most number of rules with flagged regions is over $\frac{2}{3}$).

Finally, when under the previous conditions there is either a lack of satisfied matches for a given testing observation or an unclear preference towards a single class (two or more classes with similar strength), the significance is irrelevant since the final decision is unclear. In this context, accuracy should be optimized (in detriment of significance) by: 1) decreasing the matching threshold (VI-1.8) and/or 2) combining this output with the output of other classifiers (according to *Section VI-3.2*).

6.2 Results and Discussion

Results are organized as follows. First, we measure how training and testing functions can affect the performance of the proposed methods. Second, we extend the results from previous chapter by effectively measuring the effects of the proposed principles on their performance (mean and variability components). Third, we further analyze the over- and underfitting propensity of classifiers by studying their bias and variance. The proposed principles were accommodated in FleSBiC and the algorithms were run with an Intel Core i5 2.80GHz with 6GB of RAM.

Introductory Note. Although the provided results in this section are self-explainable, we suggest the analysis of the experimental analyzes undertaken in the previous chapter (*Section VI-5.4*) since they provide a basis to understand how significance criteria shapes the behavior (and impacts the performance) of classifiers.

Impact of Training and Testing Assumptions. Figure 6.2 illustrates how training and testing decisions can impact performance by promoting susceptibility to false discoveries. For this analysis, we applied the enhanced FleSBiC on the colon dataset (high-dimensional gene expression samples to study cancer). For the analysis of training options, we applied FleSBiC with a pruning function that selects a parameterizable number of regions per class (varying from 1 to 20 in the provided analysis). Two major observations can be observed. First, a decrease in the number of regions highly degrades performance since there is an inappropriate coverage of the search space. This problem is particularly critical for FleSBiC since we allow the learned regions to be supported by a subset of overall observations. Second, the increase on the number of regions is associated with an increase in accuracy due to a lower risk of underfitting. When considering more than 10 regions, the discriminative power of the additional regions becomes considerably weak. However, this does not degrade the accuracy of FleSBic the applied integrative score adequately measures this issue. As such, FleSBiC does not incur in the risk of overfitting the input data.

For the analysis of testing decisions, we applied FleSBiC with a noise-tolerant threshold for matching new observations against the learned regions (the threshold was varied from 20% to 100%). We can observe that very relaxed matchings do not allow to understand whether an observation is described by a region, while very tight matchings (including perfect matchings when the threshold is 100%) can lead to the missing of important matchings due to intolerance to noise. These two observations are respectively associated with the risks of making false positive and false negative decisions.

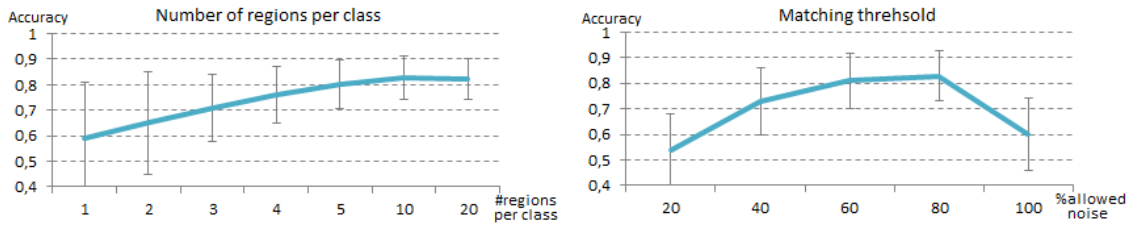


Figure 6.2: Accuracy of FleSBiC for the colon dataset when parameterized with a fixed number of regions per class and varying relaxations for matching observations.

True Performance. An inherent challenge associated with some of the conducted analyzes in the previous chapter is the inherent high variability of performance due to the high-dimensionality and low size of the input data (see Figure VI-5.7). As such, it is hard to understand which classifiers perform better. Nevertheless, for classifiers with probabilistic outputs, such as FleSBiC (able to provide the strength per class, as well as indicators of statistical significance), the analysis of their performance can be made directly from the probabilistic outputs instead of simply accounting for right-or-wrong decisions. This analysis is provided in Figure 6.3 for three associative classifiers – an early variant of FleBiC (only able to focus on few regions of the data space), FleBiC and FleSBic – applied over colon and lymphoma datasets. Interestingly, this analysis appears to better reveal their true error since it enhances differences on their performance. Illustrating, given the output $\{c_1{=}0.4,c_2{=}0.6\}$ for a testing observation. Under a smooth metric, the classifier contributes with a 0.6 error estimate (instead of 1) when c_2 is the true label (penalizing

accuracy), while is able to contribute with 0.4 when c_1 is the true label (benefiting accuracy).

Three major implications are derived from this analysis. First, the relevance of the proposed principles throughout this and previous chapters (superiority of FleBiC and FleSBiC can be tested with high confidence). Second, their role to improve accuracy (average error) and also reduce error variability. Finally, the importance of using smoothing factors whenever possible to gain a more precise measure of the true error.

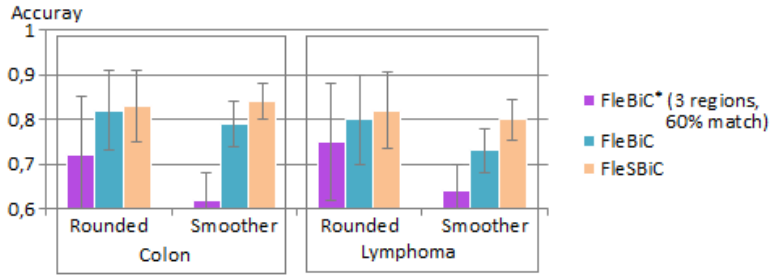


Figure 6.3: New accuracy view of FleBiC and FleSBiC based on smoothing factors to better assess their behavior over high-dimensional data (colon and lymphoma datasets).

Generalization Error (Hypothesis Testing). We conducted an analysis of the bias and variance components of the error associated with the previous results in this chapter. This was done by performing a 10-fold cross-validation and generating for each fold 100 samples using bootstrap replacement. We observed that the bias component is slightly higher than variance across settings. Understandably, a higher bias is expectable since local classifiers are focused on specific regions of the data space, and thus can miss relevant relations (underfitting propensity). The observed variance, a result from modeling the random noise in the training data (rather the intended regularities), is an inevitable result of the low number of observations of the target labeled expression data.

Learning decisions that lead to a better coverage of the data space (such as illustrated in Figure 6.2 when FleSBiC is parameterized with a large number of regions) is associated with a decrease in the bias component without affecting the variance component, thus creating an optimum balance. This balance is implicitly associated with a minimized propensity towards overfitting and underfitting. This observation validates the underlying premise of our work. In this context, we consider the study of the capacity of classifiers according to the principles proposed throughout this book to be the top priority for future work.

6.3 Conclusion

This chapter extends the contributions of the previous chapter in order to promote an adequate learning of classification models from high-dimensional data based on their guarantees of statistical significance. This is done by interpreting the overfitting and underfitting propensity of classifiers as a direct consequence of the impact that their learning functions may have on the number of false positives (included yet non-significant regions) and false negatives (neglected yet significant regions), respectively.

In this context, we explored how certain learning criteria (such as composition and matching criteria) can contribute to either an increase of false positives or false negatives, and how this impact can be measured. We further extended the assessment of the significance of a decision rule (from previous chapter) towards classification decisions, proposing new scores to support real-world decisions. Finally, we enhanced classifiers with principles to guarantee an adequate optimization of accuracy whenever significance guarantees cannot be provided.

The collected empirical evidence from real data supports the relevance of the proposed statistical view to look at classifiers. We also identify improvements in performance from the enhanced classifiers. This observation appears to be associated with their inherent ability to generalize with a minimized risk of underfitting the input data, creating an optimum balance between bias and variance components of error. As such, these contributions provide an unprecedented way to learn local classifiers with transparently assessed risks of making a false positive and/or false negative decision.

7

Multi-period Classification for Predictive Tasks

As the majority of real-world decisions change over time, classifying an attribute of interest across different time periods becomes increasingly important. Tackling this problem, referred to as *multi-period classification*, is critical to answer real-world tasks, such as the biomedical prognostics, risk anticipation, administrative planning tasks or the prediction of the evolving state of economic and geophysical systems. This chapter aims to propose adequate learners (able to accommodate the proposed principles throughout *Book VI*) for this end. Given a database with training observations derived (possibly structured) multi-attribute data and an attribute of interest, the target task of *multi-period classification* can be informally defined as the learning of a model to label the attribute of interest for a new observation across h>1 time periods (horizon). In other words, instead of outputting a single label, multi-period classifiers aim to learn a sequence of h>1 labels for new observations. An illustrative task is the planning of hospital resources by predicting if a patient will need a specific treatment within upcoming years.

Classic classification principles are not able to effectively solve this task since their separated application on each time period implies conditional independence among the periods under classification. Nevertheless, important contributions can be seized from related research streams. In particular, long-term prediction provides principles for capturing the temporal dependencies among the periods for the attribute under prediction [60]. However, these principles have not yet been extended towards classification. In multi-period classification tasks, unlike sequence prediction and forecasting, the multiple labels under classification are not the next occurrences of an observed sequence in the data domain. Instead, multi-period classification assumes independence between the attributes in the domain and the attribute under classification (codomain).

Two major requirements can thus be defined for the multi-period classification task. First, multi-period classifiers should model the stochastic dependencies underlying the periods under classification. Second, multi-period classifiers should be able to embed existing learning behavior (including the classifiers proposed throughout *Chapters VI-1-VI-6*) in order to be able to effectively deal with distinct data structures and the specificities of real-world domains. The core contribution from this paper is, thus, to study the performance of multi-period classifiers able to answer these two requirements.

This work motivates the need for multi-period classifiers, and proposes a method, Cluster-based Multi-Period Classification (CMPC), that preserves local dependencies across the periods under classification, while make use of the learning behavior of traditional classifiers. Six major contributions are provided:

- structural view on the multi-period classification task, applicability, requirements, and relation with other streams of research;
- evaluation framework for multi-period classification, including: 1) adequate loss functions for nominal and ordinal labels, 2) extended confusion matrices and derived performance views, and 3) advanced loss functions to smooth temporal misalignments and assess error propagation along the horizon;
- hybrid single-output method able to minimize error propagation yet model dependencies between periods;
- new class of multiple-output methods based on clustering (CMPC) able to: adequately reduce and recover the space of sequences, be applied on top of any single-label classifier, and address the limitations of (single-

ouptut/combinatorial/circuitry) alternatives adapted from long-term prediction;

- two variants of the multiple-output method: 1) variant based on the segmentation of the horizon of prediction to minimize the flexibility issues of the previous multiple-output method; and 2) variant based on moving sliding windows to guarantee a more accurate modeling of the true stochastic dependencies between the periods under prediction;
- dynamically parameterizable behavior of the proposed methods (including the number and properties of clusters and segments) to minimize the number of false positive and false negatives.

Evaluation against real-world datasets provides evidence of the relevance of multi-period classifiers, and shows the superior performance of the CMPC method against peer methods adapted from long-term prediction for multi-period tasks with a high number of periods. We also show additional properties of interest associated with the cluster-centric behavior of the CMPC method, including its propensity to deal with non-trivial combinations of labels and an increased sensitivity towards less frequent labels without overfitting risks.

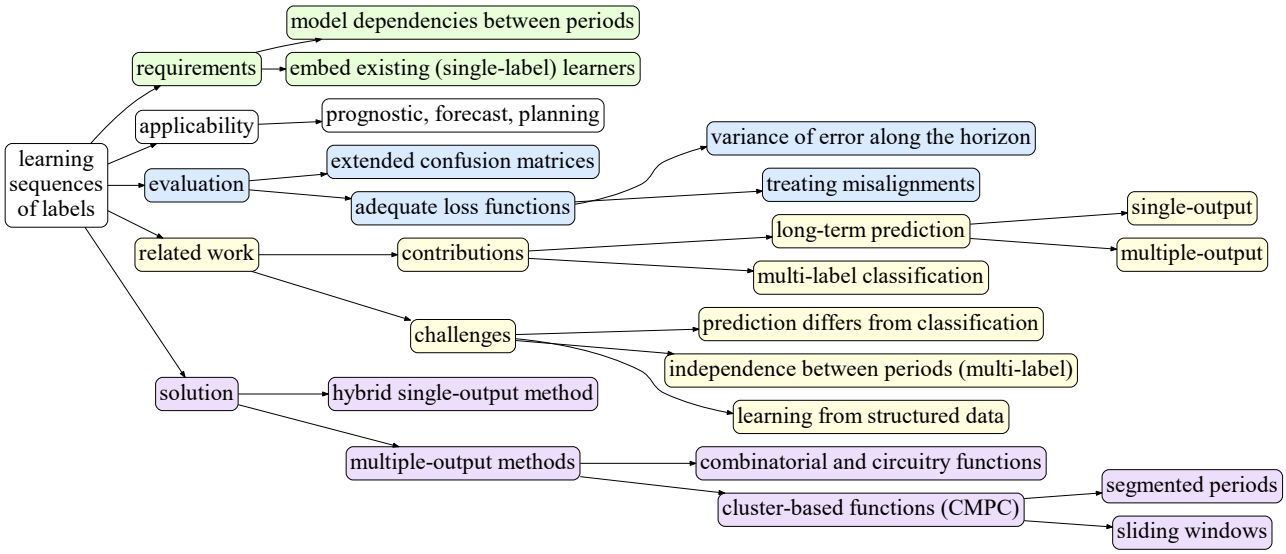


Figure 7.1: Requirements, challenges and proposed contributions for multi-period classification.

Figure 7.1 provides an overview of the challenges and proposed contributions. Accordingly, this chapter is structured as follows. *Section 7.1* formalizes the multi-period classification task and lists its major applications. *Section 7.2* surveys the contributions and limitations from related research. *Section 7.3* describes the solution space, including the CMPC approach to answer the target task. *Section 7.4* describes new evaluation metrics to adequately assess the contributions in this field. Finally, the performance of CMPC is evaluated and discussed in *Section 7.5*, and the resulting implications and possible directions for future work covered in *Section 7.6*.

7.1 Background

In what follows, *Section 7.1.1* provides a formal view of the task and its requirements, and *Section 7.1.2* motivates its increasing need to answer a wide-range of prominent real-world problems.

7.1.1 Formalization

Multi-period classification, the task of learning a mapping model to estimate a sequence of labels for an unlabeled observation from a training set of labeled observations (Def.7.1, should not be mingled with:

- *prediction* tasks, the supervised estimation of the next occurrences of a (univariate) sequence based on a set of observed sequences;

- *forecasting* tasks, the unsupervised estimation of upcoming events for a time sequence;
- *regression* tasks, the supervised learning of parametric models to estimate a numeric attribute (C=$\mathbb{R}$).

Def. 7.1 Let an univariate sequence be $\mathbf{c} = \{c_i \mid i = 1,..,h\} \in \mathbb{T}^h$, where $\mathbb{T}^h$ is the set of all sequences with $h \in \mathbb{N}$ length and elements $c_i \in C$. Consider a dataset $\mathbf{A}$ with n labeled observations $(\mathbf{x}_i, \mathbf{c}_i)$, where $\mathbf{x}_i \in \mathcal{X}$ is an observation from a (possibly structured) domain, and $\mathbf{c}_i \in \mathbb{T}^h$ is a sequence with $h > 1$ labels (referred as the *horizon of prediction*) from an ordinal or nominal alphabet C.

Given $\mathbf{A}$, the **multi-period classification** task aims to learn the mapping model $M : \mathcal{X} \rightarrow \mathbb{T}^h$ for labeling new tuples $(\mathbf{x}_{new}, \mathbf{c}_{new})$, where $\mathbf{c}_{new}$ is unknown, that is, $\hat{\mathbf{c}}_{new}$=$M(\mathbf{x}_{new})$, where $\hat{\mathbf{c}}_{new}$ is the sequence of estimated labels, $\mathbf{x}_{new} \in \mathcal{X}$, $\hat{\mathbf{c}}_{new} \in \mathbb{T}^h$ and $h \in \mathbb{N} \wedge h > 1$.

Multi-period classification can be consistently formulated independently of the properties of the data domain $\mathcal{A}$, which can be tabular (where an observation is associated with a set of features) or structured (where an observation is possibly associated with a multivariate time series or with a multi-set of events). Illustrating, consider ($\mathbf{x}_i$,$\mathbf{c}_i$=<good,risk,hospitalization>) to be an observation associated with a patient from a healthcare data domain where the attributes are either event-sets or simple features (such as age) and $\mathbf{c}_i$ describes the categoric health states across sequent periods to study the need for upcoming interventions.

Note that we do not impose an uniform temporal range for the periods under classification since $\mathbf{c}$ labels are simply described by a time series. These periods may: 1) have different durations (**non-uniform condition**) and 2) allow for temporal gaps (**non-convex condition**). Multi-period classification can be additionally applied to learn the order of the upcoming h events.

In this context, two major requirements can thus be defined for the multi-period classification task:

- **R1**: *Modeling stochastic dependencies* underlying the sequence under classification;
- **R2**: *Embedding existing learning functions* to adequately learn from domains with varying properties.

7.1.2 Applications

The prediction of the evolving state of living, geophysical, economic and societal systems is referred as one of the ten most critical data mining challenges for this decade [693]. The core applications of multi-period classification can be synthesized according to three key areas.

First, the prediction of both local and global *emerging trends* for: *i)* catastrophe anticipation of epidemics or environmental changes that may lead to hurricanes or seismic activity; and *ii)* the assessment of key changes in human health (both from a clinical, psychophysiological and biological perspective) and behavior (based on temporal data derived from social networks and web usage logs).

Second, the support to *personalized decisions*, such as the prognosis of a patient's health condition over time based on its health-records or the estimation of upcoming user behavior based on its monitored actions and profile. Complementarily, the identification of the regularities discriminating a specific sequence of labels (such as biomedical markers determining disease progression or social markers anticipating key behavior of interest) is increasingly relevant.

Finally, the support of *planning tasks* for almost every system that records relevant events. An application is the planning of physical resources (increasingly necessary for healthcare and commerce value-systems). Another option is the estimation of critical variables to uphold transparent budgets and decisions in both private industries and public sectors as education and social security.

7.2 Contributions and Limitations from Related Work

Despite the large research on learning classifies from a wide-variety of data structures (see previous chapters), less attention has been given to the problem of learning sequences of classes (symbolic time series). In fact, despite

the relevance of multi-period classification, the first dedicated attempts to formalize the problem and systematize its requirements were only recently proposed by the authors [317, 306]. Below, we overview the contributions of two related research streams – multi-label classification and long-term prediction – and synthesize why they are insufficient to answer the multi-period classification task.

Multi-label Classification. A task that allows the classification of multiple classes is multi-label learning. Given a training set of observations with h disjoint nominal labels, the task of multi-label classification is to learn a mapping model to predict h labels for a new observation, where $h > 1$.

The common option to deal with multiple classes is to map this task into a set of single-label classification tasks by performing data transformations [636]. The simplest mapping is to learn h classifiers (one for each c_i period), with the output being the union of the learned labels. Alternative transformations consider the definition of coverage-based classifiers or the combinatorial view of multiple classes as a new class. Classifiers, such as decision trees with modified entropies or lazy learners with label-ranking probabilities [707], have been proposed for an improved performance in multi-label settings.

However, since multi-label learning was developed for categorization and multi-parameter diagnosis, it does not consider conditional dependencies among the target labels (7.1). It satisfies **R2** but it does not satisfy **R1**.

$$P(\mathbf{c} \mid \mathbf{x}) = P(\{c_1, ..., c_h\} \mid \mathbf{x}) = \Pi_{i=1}^{h} P(c_i \mid \mathbf{x}) \tag{7.1}$$

Long-term Prediction. A related research stream with key principles for learning series of values is long-term prediction. Given a training dataset with sequences with $m + h$ elements, the long-term prediction task aims to learn a model, $M : \mathbb{T}^m \rightarrow \mathbb{T}^h$, to predict the $h > 1$ next elements of a m-length sequence.

Long-term prediction commonly follows single-output mappings. Single-output methods rely on the multiple application of one-step-ahead predictors by learning models that either use or discard estimations across the periods under prediction. One-step-ahead predictors have been intensively researched for both *numeric* sequences and, more interestingly for the target multi-period task, *categoric* sequences [467]. In the latter case, one-step-ahead predictors either rely on *generative models*, such as dynamic Bayesian networks, Markov chains, time-delay neural networks (NNs) and recurrent NNs, or on supervised *predictive rules* for constraining future events [398]. There are two major types of single-output methods: iterative and direct. In *iterative* single-output methods [85], a h-step-ahead predictor is defined by iterating, h times, the one-step-ahead predictor. In each iteration the estimated values are used as inputs for the next iteration. This has an evident negative impact in terms of error propagation [594]. *Direct* single-output methods perform the h-step-ahead prediction by learning h models, each returning a estimate that does not depend on previous estimations. Although not prone to error accumulation, direct methods are associated with higher functional complexity (as defined in [86]) to model the stochastic dependencies between the observed sequences and arbitrary distant elements as they are not able to consider the underlying dependencies among the predicted variables [86]. DirRec [595] is a *hybrid* method that offers a compromise between the behavior of direct and iterative methods. Extensions have been proposed to deal with non-stationary sequences using Gaussian process models with a modified covariance function [94], to learn multi-modal distributions underlying the sequence for predictions with more delineated changes in values [419] and to deal with cyclic behavior using modular NN architectures [62].

Contrasting with single-output methods, Multiple-Input *Multiple-Output* (MIMO) methods learn one model to classify all periods at a time. The goal is to preserve the stochastic dependencies for a reduced bias. However, these methods reduce the flexibility of single-output approaches [61]. To avoid this, Taieb et al. [613] proposes intermediate configurations by decomposing the original task into $k{=}h/s$ tasks, where s is the number of time periods per task predicted at a time. This approach, called *Multiple-Input Several Multiple-Outputs* (MISMO), trades off the property of preserving the stochastic dependency among periods with a greater flexibility of the predictor.

Experiments show that the choice of s strongly varies according to the input data, with s=1 (direct method) and s=h (MIMO method) being good performers in less than 20% of the cases [60]. Multiple-output predictors are less frequent than single-output predictors since they require changes on the learning kernels. Lazy learners have been proposed with discrepancy assessments and averaging strategies to cope with medium-to-large horizons of prediction [86]. In order to avoid local minima problems, Ji et al. [355] extended least-squares support vector machines as a MIMO method.

Basics 7.1 Illustrating the behavior of long-term predictors

An illustrative prediction of the 4^{th} period for a testing observation $\mathbf{x}_i$ would be: $\{\mathbf{x}_i\}\rightarrow c_4$ (direct single-output), $\{\mathbf{x}_i, \hat{c}_1, \hat{c}_2, \hat{c}_3\}\rightarrow c_4$ (iterative single-output), $\{\mathbf{x}_i\}\rightarrow\{c_1, c_2, c_3, c_4\}$ (multiple-output with s=4) and $\{\mathbf{x}_i, \hat{c}_1, \hat{c}_2\} \rightarrow \{c_3, c_4\}$ (multiple-output with s=2).

Despite the relevance of these contributions, long-term predictors cannot be straightforwardly used for multi-period classification. First, in multi-period classification tasks, the sequence of labels $\mathbf{c}$ is not necessarily a follow-up of a sequence in the data domain. Therefore, the common parametric functions used in long-term prediction are not well suited for this task. Second, the existing MIMO/MISMO models for capturing local dependencies across the periods under classification require a case-by-case adaptation of the learning methods, thus not satisfying **R1**.

7.3 Solution: Methods for Multi-period Classification

Despite the previously proposed contributions throughout this book to learn classifiers from (possibly structured) temporal data, there are not yet solid principles on how to effectively extend them to label an attribute of interest across sequent periods. For this aim, the methods proposed below are designed to satisfy the **R2** requirement, thus preserving the original core learning behavior by by relying on one or more instantiations of a single-label classifier. Figure 7.2 provides a taxonomy of the covered multi-period methods in this section. As illustrated, the proposed multi-period solutions are independent from the learning specificities of the chosen classifier. In this way, they can be applied over distinct data structures.

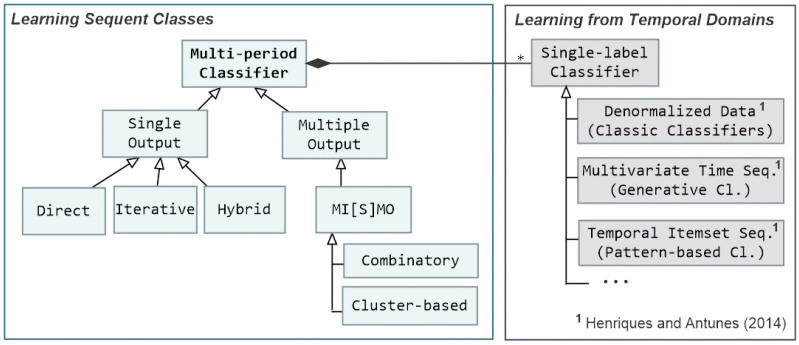


Figure 7.2: Multi-period classifiers seen as an extension of single-label learners. Three single-output strategies (iterative, direct and hybrid) and two multiple-output strategies (combinatorial and cluster-based) are provided for the task of multi-period classification and independent from the core learning behavior.

Def. 7.2 Consider a dataset $\mathbf{A}$ with n labeled observations $(\mathbf{x}, c_1, .., c_h)$.
Single-output approaches to multi-period classification label c_i: 1) directly, $M_{i\in\{1,..,h\}}(\mathbf{x})$, where h is the horizon of prediction and M is classification model; or 2) iteratively, $\{M_{i=1}(\mathbf{x}), M_{i\in\{2,..,h\}}(\hat{c}_i, .., \hat{c}_1, \mathbf{x})\}$.
Multiple-output approaches to multi-period classification label h=ks periods (within k steps of s periods each), $\{\hat{c}_{ps}, .., \hat{c}_{(p-1)s+1}\} = M_p(\mathbf{x})$, with $p \in \{1, .., k\}$, where s is the model variance. The model variance, s, constrains the perseveration of stochastic dependencies of the sequence: null if s=1 and maximal if s=h.

Def.7.2 maps the single-output and multiple-output methods from long-term prediction into the context of multi-period classification. These strategies simply rely on the multiple instantiation-and-learning of single-label classifiers to deliver the target multi-period behavior. However, single-output solutions either suffer from error propagation or from heightened functional complexity. To address this problem, *Section 7.3.1* proposes an hybrid

single-output method to balance and minimize the errors of peer single-output methods.

Nevertheless, a shared problem by all single-output methods is that, since each period is classified independently, they fail to capture emerging local dependencies among the periods under classification. Additionally, existing multiple-output methods proposed in long-term prediction require a case-by-case adaptation of existing learning settings. This forbids the use of the previously proposed single-label classifiers. In this context, *Section 7.3.2* proposes a new multi-output approach able to surpass these challenges, and *Section 7.3.3* extends this approach towards a MISMO variant.

7.3.1 Hybrid Single-Output Classifiers

To trade-off the behavior of iterative and direct strategies, simple hybrid methods that learn both direct and iterative classifiers for each period and select the classifier with best performance can be adopted. Direct and iterative models are learned using a training dataset and evaluated either over the same training observations or over a dedicated set of observations for this purpose. Although this hybrid setting can improve the multi-period performance, it obligates to either use all past predictions or none of them for a specific period under classification, depending on whether an iterative or direct strategy is selected for that period. Understandably, this is only useful when there are alternate groups of sequent periods with strong dependencies (preference towards iterative approaches) or with very low structural dependencies (preference towards direct approaches). However, in real-world data there is no such clear boundary on whether to use all or none of the past period predictions.

Therefore, a new hybrid strategy is proposed. This strategy removes the periods associated with the selection of a direct model from upcoming period estimations. Algorithm 11 describes this strategy.

Algorithm 11: Hybrid single-output method to multi-period classification with removal of direct estimations (single-output decisions tested on a validation set).

Notation: c^i_j denotes the i^{th} period of sequence $\mathbf{c}_j$ labeled with observation $\mathbf{x}_j$

Method: ***BuildHybridClassifier***
Input: $\{\mathbf{a}_1{=}(\mathbf{x}_1,\mathbf{c}_1),..,\mathbf{a}_n{=}(\mathbf{x}_n,\mathbf{c}_n)\}$ *(observations)*, $M' : \mathcal{X} \rightarrow C$ *(single-label learner)*
Output: $M = \{M'_1,..,M'_h\}$ *(hybrid multi-period classifier* $M : \mathcal{X} \rightarrow \mathbb{N}^h$*)*
$\{\mathbf{c}^{past}_1,..,\mathbf{c}^{past}_n\} \leftarrow \{\emptyset,..,\emptyset\}$;
foreach $i \leftarrow 1$ **to** h **do**
  Part I: learning and assessment of direct and iterative models for the i period
  foreach $j \leftarrow 1$ **to** n **do** $\mathbf{a}'_j \leftarrow (\mathbf{x}_j \cdot \mathbf{c}^{past}_j, c^i_j)$; *(add past predictions on* $\mathcal{X}$*)*
  $M^{direct}_i \leftarrow$ buildDirectClassifier(M', trainPortion($\{(\mathbf{x}_1,c^i_1),..,(\mathbf{x}_n,c^i_n)\}$));
  $M^{iterative}_i \leftarrow$ buildIterativeClassifier(M', trainPortion($\{\mathbf{a}'_1,..,\mathbf{a}'_n\}$));
  correctDirect $\leftarrow 0$; correctIterative $\leftarrow 0$;
  foreach $j \in$ testIndexes($\{\mathbf{a}_1,..,\mathbf{a}_n\}$) **do**
    if $M^{direct}_i(\mathbf{x}_j){=}c^i_j$ **then** correctDirect $\leftarrow$ correctDirect + 1;
    if $M^{iterative}_i(\mathbf{x}_j \cdot \mathbf{c}^{past}_j){=}c^i_j$ **then** correctIterative $\leftarrow$ correctIterative + 1;
  Part II: remove periods associated when a direct-method is preferred
  if correctIterative > correctDirect **then foreach** $j \leftarrow 1$ **to** n **do** $\mathbf{c}^{past}_j \leftarrow \mathbf{c}^{past}_j \cdot c^i_j$; *(concatenate i-period label)*;
  $M'_i \leftarrow$ buildClassifier($M^{iterative}_i$,$\{\mathbf{a}'_1,..,\mathbf{a}'_n\}$);

Method: ***ClassifyObservation***
Input: $\mathbf{a}_{new}{=}(\mathbf{x}_{new},\emptyset)$ *(observation with unobserved sequence)*, $M{=}\{M'_1,..,M'_h\}$ *(multi-period classifier)*
Output: $\hat{\mathbf{c}}_{new}$ *(sequence of classified periods, initialized with* $\hat{\mathbf{c}}_{new} \leftarrow \emptyset$*)*
$\hat{\mathbf{c}}_{past} \leftarrow \emptyset$; *(time series with past iterative predictions)*
foreach $i \leftarrow 1$ **to** h **do**
  $\hat{c}^i_{new} \leftarrow M'_i(\mathbf{x}_{new} \cdot \hat{\mathbf{c}}_{past})$;
  if *isIterative*(M'_i) **then** $\hat{\mathbf{c}}_{past} \leftarrow \hat{\mathbf{c}}_{past} \cdot \hat{c}^i_{new}$; *(concatenate iterative prediction)*;
  $\hat{\mathbf{c}}_{new} \leftarrow \hat{\mathbf{c}}_{new} \cdot \hat{c}^i_{new}$ *(add most probable label)*

7.3.2 Cluster-based Multi-Period Classification

Multi-output models return a estimation of the $\mathbb{T}^h \mid \mathcal{X}$ distribution by taking into account the dependencies between all periods of $\mathbb{T}^h$ within a single step. The few algorithms that learn these models suffer from a critical drawback –

they are not independent from the core learning. This results from the need to adapt the behavior of the single-label classifier to perform multi-period tasks. Some illustrative single-label classifiers with modified kernel functions are local learners [86], SVMs [355] and NNs [62].

7.3.2.1 Combinatorial Behavior

The simplest way to deal with the classification of multiple values at a time is to view them as a single value. A naïve strategy is to rely on the enumeration of all possible sequences of labels, which results in $\binom{h}{|\mathcal{A}|}$ classes.

Understandably, the great limitation is that this number is usually very high either for large horizons or for lengthy alphabets, which easily leads to a training dataset with a small number of observations per class.

Basics 7.2 Circuitry to reduce exponential combinations of labels

A strategy to reduce the possible number of classes is to learn coding-and-decoding functions able to describe combinatorions of labels among the periods of the horizon of prediction. For simplicity sake, let us assume a binary circuit for binary classes. Figure 7.3 illustrates a circuit where an horizon of prediction is characterized by contiguous local dependencies. This approach suffers from the problem of requiring a pre-settled designed circuit. To solve this problem, two techniques can be employed. First, the definition and use of template circuits, as customizable versions of the illustrated one, sensitive to different parameters as the horizon length and the window of dependencies. Second, the unsupervised learning of these circuits using the training dataset, which can rely on techniques available from research on circuits and architecture self-learning [199, 603]. Binary circuits can be extended to support categoric classes [491].

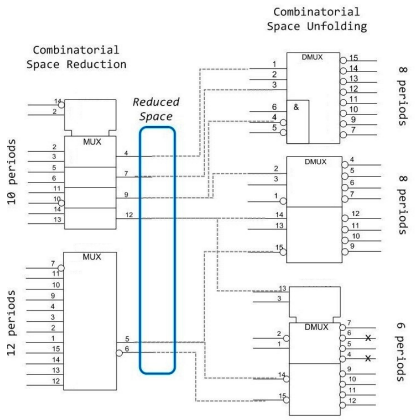


Figure 7.3: Illustrative circuit to reduce the combinatorial space for multi-period classification

In particular, we can generalize the notion of circuits towards complex coding-and-decoding functions. In fact, these functions can be as complex as a classification model since they similarly relate a set of domain values with a particular class. The key difference is that their architecture must be provided apriori or learned in an unsupervised fashion. The drawback of circuits and peer functions resides on the need to provide customizable templates to guide their learning.

7.3.2.2 Cluster-based Behavior

To overcome the problem of combinatorial and circuit-based strategies, we propose a robust alternative: Cluster-based Multi-Period Classification (CMPC). CMPC methods identify and abstract relevant sequential behaviors across the horizon of prediction to perform multi-period classification.

CMPC can be described according to five simple steps. First, the labels across the horizon for each one of the training observations is used as the input for a clustering method. Second, the learned clusters are used to replace the h labels from each training observation by the respective cluster (now seen as the class). Third, a target single-label classifier is learned from this new set of training observations. Fourth, testing observations are classified with a particular cluster. Finally, the centroid of the clusters is recovered and adopted as the target prediction $\hat{\mathbf{c}}$. These steps are described in Algorithm 12, which receives a dataset and a single-label classifier as input and outputs the multi-period model M.

Figure 7.4 illustrates the steps of CMPC for a setting where labels correspond to a set of ordinal values. In this illustrative setting, we can observe that clustering is able to reduce the search space by mapping similar $\mathbf{c}$ sequences

(such as the third and fourth sequences) or sub-sequences (such as the two first partitions of the first sequence) as a single cluster characterized by the centroid values. Furthermore, and unlike existing multi-output approaches, independence from the learning setting is guaranteed since any single-label classifier can be consider within CMPC without the need of being adapted.

Algorithm 12: Cluster-based Multi-Period Classification: MIMO variant.

Method: ***BuildMIMOClusterClassifier***
Input: $\{\mathbf{a}_1=(\mathbf{x}_1,\mathbf{c}_1),..,\mathbf{a}_n=(\mathbf{x}_n,\mathbf{c}_n)\}$ *(observations)*, $M' : \mathcal{X} \rightarrow C$ *(single-label learner)*
Variables: Σ *(cluster's identifiers)*,
$Cod : \mathbb{T}^h \rightarrow \Sigma$ *(parameterized clustering method)*,
$Dec : \Sigma \rightarrow \mathbb{T}^h$ *(cluster decoder)*
Output: M=(M',Dec) *(multi-period classifier,* $M : \mathcal{X} \rightarrow \mathbb{T}^h$*, is a pair* ($M'$,$Dec$))
$\Sigma \leftarrow$ setClusters(Cod,$\{\mathbf{c}_1,..,\mathbf{c}_n\}$,$min(|C| \times \sqrt{h}, \frac{n}{5})$,0.2); *(Elbow local optimization)*
$Cod \leftarrow$ learnClusteringModel(Cod,$\{\mathbf{c}_1,..,\mathbf{c}_n\}$,$\Sigma$);
$Dec \leftarrow$ buildDecoder(Cod); *(centroids of discrete model or the means of multi-Normal distribution model)*
foreach $j \leftarrow 1$ **to** n **do**
  $\mathbf{a}'_j \leftarrow (\mathbf{x}_j, Cod(\mathbf{c}_j))$; *(replaces the sequence by the learned cluster)*
$M' \leftarrow$ learnClassificationModel($\{\mathbf{a}'_1,..,\mathbf{a}'_m\}$);
$M \leftarrow (M',Dec)$;

Method: ***ClassifyObservation***
Input: $\mathbf{a}_{new}$=($\mathbf{x}_{new}$,$\emptyset$), M=(M',Dec)
Output: $\hat{\mathbf{c}}_{new}$ *(sequence of estimated labels)*
$\sigma \leftarrow M'(\mathbf{x}_{new})$; *(estimate cluster* $\sigma \in \Sigma$*)*
$\hat{\mathbf{c}}_{new} \leftarrow Dec(\sigma)$; *(derive sequence of labels)*

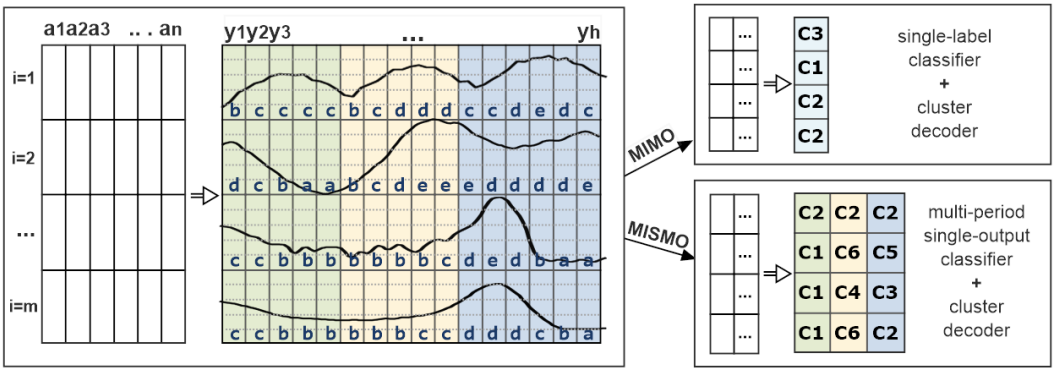


Figure 7.4: Illustrative application of cluster-based multiple-output methods. The two variants, CMPC-MIMO and CMPC-MISMO (with s=4), are applied to classify an ordinal attribute with $|C|$=5 labels for h=16 periods.

7.3.2.3 Advanced Aspects

Beyond the classifier choice, the clustering method can be either input or dynamically parameterized based on the number of periods, labels and observations, and on the diversity and representativity of observed sequential behavior. Three variables are considered to define adequate clustering methods:

- *Algorithmic choice.* Sequence clustering is the default option in CMPC to guarantee its correct ability to model contiguity of labels and to efficiently deal with a high number of periods in the presence of large horizons of prediction (high-dimensional symbolic time series) [546].
  The clustering paradigm – either greedy (sequential k-Means [299]) or stochastic (expectation-maximization [93]) – is dynamically selected depending on whether the distribution of observations closely follows a Gaussian distribution.
  Finally, a set of wrappers can be optionally allowed, including density-based wrappers to obtain a new centroid criterion for the alleviation of the problems related to greedy methods [377], and hierarchical-based wrappers [667] to consider alternative combinatorial space reductions;
- *Number of clusters.* The number of clusters is dynamically defined as a function of the context, based on $|C|$, h, n, and the dispersion of the observed sequences of labels $\{\mathbf{c}_1,..,\mathbf{c}_n\}$. The number of clusters should

not be too low in order to capture the specificities of different sequential behavior and to avoid time series' centroids without delineated differences. Additionally, an excessive number of clusters result in models with low number of observations per class. Although the number of clusters can be fixed using well-known techniques as the k-means++ method, silhouettes or iterative validation of the clustering error [561, 605]; these techniques are not prepared to guarantee that the fixed number of clusters provides an adequate basis for classification.

For this aim, CMPC sets the initial number of clusters to $min(|C|\times\sqrt{h}, \frac{n}{5})$ and then dynamically optimizes this number using the Elbow approach [625] under the guarantee that the number of clusters does not deviates more than 20% from the initial number. $|C|\times\sqrt{h}$ guarantees that distinct sequential behavior of interest is selected, while $\frac{n}{5}$ guarantees that, at least, approximately 5 observations are considered for each class (cluster) in order to minimize the risk of overfitting associated with the learning task;

- *Distance Metrics.* Alternative sequence similarity functions, ranging from simple distance metrics as Euclidean (default option) to more advanced ones [176] can be adopted.

  Complementary, the centroid metric can be computed using the median labels (default option), the ward, and other metrics based on operators as neighbor joining.

7.3.3 Extension of CPMC to Capture Local Temporal Dependencies

The drawback of the CMPC under a MIMO variant is its weak flexibility when compared with single-output methods. In particular, for a high number of periods, increasing the number of clusters does not necessarily improve accuracy since the number of training observations per class decreases. To tackle this problem, we propose the CMPC extension towards a MISMO variant.

By segmenting the periods under prediction, the problem of losing flexibility under lengthy horizons is alleviated. This extension can be implemented using a direct or iterative (multiple-output) strategy, where the cluster-based model is applied for each segment along the horizon of prediction. This strategy trades off the preservation of stochastic dependencies among future values with a greater flexibility of the classifier.

The increased MISMO adaptivity comes at the cost of an additional parameter to define the number of periods per segment, the model variance s. Model variance can be fixed using a sensitivity analysis, $s=argmax(\{acc_h, acc_{h/2}, acc_{h/3}, .., acc_{h/h}\})$, or as function of both the dimensionality h and the stochastic properties of $\{\mathbf{c}_1, .., \mathbf{c}_m\}$ training sequences. If these sequences show periodicities or local stationarity, the length of these local segments can be used to estimate s. However, this option is not applicable to non-stationary sequences.

To address this challenge, CMPC relies on a different methodology for the estimation of s. For a given s', consider the clustering solution learned from the observed h/s' segments for each observation of the training dataset using $|C|\times\sqrt{s'}$ clusters. By starting with $s'=h$ and iteratively incrementing the number of segments, new clustering solutions are computed and evaluated until the normalized clustering error (within-cluster sum of squares) no longer decreases. The resulting number of segments defines the model variance, $s=s'$.

The MISMO variant of CMPC differs from the original CMPC in three aspects. First, the periods under prediction are segmented for each observation using the s-variance. The clusters are learned using the observed behavior for all the segments and every segment is labeled with the respective cluster. Second, instead of learning a single classifier, one classifier is learned for each one of the segments following either an iterative or direct setting. Third, the predicted clusters for each segment are replaced by prototype sequences given by a centroid criteria and, finally, these sequences are concatenated for each observation to compose the estimated sequence of labels. The CMPC-MISMO variant with iterative behavior is described in Algorithm 13 and illustrated in Figure 7.4.

An apparent concern related to CMPC-MISMO approaches is the loss of relevant sequential behavior that appears when considering the whole horizon but not within each segment. Empirical analyzes show, however, that the iterative inclusion of already classified segments for the classification of next segments alleviates this problem.

Algorithm 13: Cluster-based Multi-period Classification: MISMO variant.

Method: ***BuildMISMOClusterClassifier***
Input: $\{\mathbf{a}_1{=}(\mathbf{x}_1,\mathbf{y}_1),..,\mathbf{a}_n{=}(\mathbf{x}_n,\mathbf{y}_n)\}$ *(observations)*, s *(model variance)*, M' *(single-label learner)*
Variables: Σ *(cluster's identifiers)*, B *(training sequence segments)*, q *(number of segments)*, $Dec : \Sigma \to \mathbb{T}^s$ *(cluster decoder)*, $Cod : \mathbb{T}^s \to \Sigma$ *(parameterized clustering method)*
Output: $M{=}(\{M'_1,..,M'_q\},Dec)$ *(multi-period classifier,* $M : \mathcal{X} \to \mathbb{T}^h$*)*
$B \leftarrow \emptyset$;
If $s \leq 0 \vee s > h$ **then** $s \leftarrow$ estimateVariance(Cod,$\{\mathbf{c}_1,..,\mathbf{c}_n\}$) *(dynamic s selection)*
$q \leftarrow h/s$;
foreach $j \leftarrow 1$ **to** n **do**
  $B \leftarrow B \cup \{\mathbf{c}_j^1,..,\mathbf{c}_j^q\}$; *(segmentation of each sequence* $\mathbf{c}_j$ *in q subsequences)*
$\Sigma \leftarrow$ setClusters(Cod,B,$min(|\mathcal{C}| \times \sqrt{s}, \frac{n}{5})$,0.2); *(Elbow local optimization)*
$Cod \leftarrow$ learnClusteringModel(Cod,B,Σ);
$Dec \leftarrow$ buildDecoder(Cod); *(centroids of discrete models or means from multi-Normal distributions)*
foreach $i \leftarrow 1$ **to** q **do**
  if *iterativeFlag* **then**
    foreach $j \leftarrow 1$ **to** n **do**
      (adds previous clusters in the domain and new cluster in the codomain)
      $\mathbf{a}'_j \leftarrow (\mathbf{x}_j \cdot \{Cod(\mathbf{c}_j^1),..,Cod(\mathbf{c}_j^{i-1})\}, Cod(\mathbf{c}_j^i))$;
  else
    foreach $j \leftarrow 1$ **to** n **do**
      $\mathbf{a}'_j \leftarrow (\mathbf{x}_j, Cod(\mathbf{c}_j^i))$; *(direct: creates observation with the learned cluster)*
  $M'_i \leftarrow$ learnClassificationModel(M',$\{\mathbf{a}'_1,..,\mathbf{a}'_n\}$);
$M \leftarrow (\{M'_1,..,M'_q\},Dec)$;

Method: ***ClassifyObservation***
Input: $\mathbf{a}_{new}{=}(\mathbf{x}_{new},\emptyset)$, $M{=}(\{M'_1,..,M'_q\},Dec)$
Output: $\hat{\mathbf{c}}_{new}$ *(sequence of estimated labels)*
$\hat{\mathbf{c}}_{new} \leftarrow \emptyset$;
foreach $i \leftarrow 1$ **to** q **do**
  $\sigma_i \leftarrow M'_i(\mathbf{x}_{new})$; *(estimate cluster* $\sigma_i \in \Sigma$*)*
  $\hat{\mathbf{c}}_{new} \leftarrow \hat{\mathbf{c}}_{new} \cdot Dec(\sigma_i)$; *(concatenate labels of segement i)*

7.3.4 CMPC with Sliding Windows

A variant of the CMPC-MISMO based on sliding-windows can be considered when the set of training observations is small or not sufficiently diverse to identify discriminative sequences of labels. The sliding-window approach can be implemented by shifting the initial s-size segment one period at a time.

However, this solution penalizes efficiency and requires the inclusion of a voting stage for periods with more than one estimated label.

Concluding Note. Cluster-based approaches for multi-period classification provide five core potentialities. First, cluster-based approaches provide an inherent simple and robust method for space reduction and recovery. Second, they address the limitations of single-output methods and provide principles to minimize the pinpointed flexibility issues of MIMO (by segmenting the horizon of prediction) and MISMO (by moving sliding windows). Third, the application of clustering guarantees an adequate number of class-conditional observations (surpassing overfitting problems) and provides an elegant way to allow the output of less-trivial combinations of labels. Fourth, the easily parameterizable behavior can be used to adjust the propensity to over- and underfitting of the CMPC learner by controlling the number of clusters and the model variance. Finally, they can be applied without the need to adapt the selected single-label classifier. Understandably, these potentialities come with the necessary cost of losing information by grouping similar sequences as single-values defined by the learned clusters.

7.4 Evaluation Metrics

Multi-period classification requires different evaluation metrics than those used in traditional (single-label) classification. As such, *Section 7.4.1* provides an extension of accuracy and confusion matrices towards multi-period contexts with nominal codomains. *Section 7.4.2* parameterizes these performance views with more adequate loss functions to deal with ordinal codomains. Finally, *Sections 7.4.3* and *7.4.4* provide complementary metrics to

correctly assess misaligned outputs, error propagation and overfitting propensity.

We propose the use of a *10-fold cross-validation* scheme to compute these metrics. Additionally, the significance of the observed differences per metric should be statistically tested using a *paired two-sample two-tailed t-test* following a *t*-Student distribution with 9 degrees of freedom and testing folds preserved across settings.

7.4.1 Baseline Evaluation

Accuracy. The accuracy of a multi-period learning model is the probability that the classifier correctly labels multiple periods for the set of testing observations (7.2).

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n} Acc_i(\mathbf{c}_i, \hat{\mathbf{c}}_i) \tag{7.2}$$

where $\mathbf{c}_i = \{c_i^1, .., c_i^h\}$ is the correct sequence of labels for i^{th} observation and $\hat{\mathbf{c}}_i = \{\hat{c}_i^1, .., \hat{c}_i^h\}$) its multi-period estimation. Multi-period accuracy Acc_j can be derived from loss functions applied along the horizon per observation. If the attribute under classification is *nominal*, the accuracy can simply be computed from the fraction of correct labels.

$$Acc_i(\mathbf{c}_i, \hat{\mathbf{c}}_i) = \frac{1}{h}\sum_{j=1}^{h}(c_i^j = \hat{c}_i^j) \tag{7.3}$$

When the periods under classification are *ordinal*, labels can be replaced by their corresponding real-values for the computation of loss functions (see *Section 7.4.2*). In multi-label learning, additional functions have been proposed to weight costs for false positives and true negatives and to detect xor differences [636]. Complementarily, other similarity functions that treat label misalignments [176] can be applied to further study the performance of multi-period classifiers performance (see *Section 7.4.4*).

Metrics from Extended Confusion Matrices. In datasets where classes are non-balanced or hold properties that turn the classification task more complex for a subset of observations, accuracy views do not suffice. Considering the healthcare case where only few patients are candidates for hospital intervention. A classifier can achieve a high accuracy by simply labeling all the periods from all patients as 'no-intervention'. Complementary views include *sensitivity*, fraction of observations with $c \in C$ label correctly identified, *specificity*, fraction of observations without c label correctly identify. F-measure can be used to trades-off sensitivity and specificity in a single metric (balanced F-Measure by default).

For the multi-period classification task, a traditional *confusion matrix* can be computed for each label and for each one of the target h periods. This solution, illustrated in Figure 7.5, has the undesirable property of not offering compact views to study performance. For instance, a sensitivity/specificity metric needs to be computed for each label and period in order to have a global view of the multi-period classifier sensitivity. Understandably, this can result in impracticable high number of metrics ($|C| \times h$).

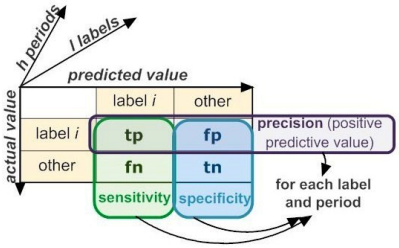


Figure 7.5: Confusion matrices in multi-period classification settings. A confusion matrix in multi-period settings is the composition of classic confusion matrices per label and period, which results in a total of $|C| \times h$ views.

To turn this assessment practicable, there is the need to collapse the high number of metrics from this setting. One option is to compute a mean across the periods to eliminate the time dimension (equations (7.4) and (7.5)).

$$Sensitivity_c = \frac{1}{h}\sum_{j=1}^{h}\frac{\Sigma_{i=1}^{n}(c = c_i^j) \wedge (c_i^j = \hat{c}_i^j)}{\Sigma_{i=1}^{n}(c = \hat{c}_i^j)} \tag{7.4}$$

$$Specificity_c = \frac{1}{h}\sum_{j=1}^{h}\frac{\Sigma_{i=1}^{n}(c \neq c_i^j) \wedge (c_i^j = \hat{c}_i^j)}{\Sigma_{i=1}^{n}(c \neq \hat{c}_i^j)} \tag{7.5}$$

where c is the class under assessment, and c_i^j and $\hat{c}_i^j$, respectively, the observed and predicted value for the j-period and i-observation.

To assess if two classifiers, M_1 and M_2, differ with regards to a specific function f, such as *sensitivity*, we propose the simple average of the absolute differences across the C labels. Given a specific function of labels f, this difference is represented as $\triangle f = \frac{1}{|C|}\Sigma_{c\in C}|f_{c|M_1} - f_{c|M_2}|$.

7.4.2 Multi-period Classification with Ordinal Labels

Multi-period accuracy Acc_i in the presence of ordinal labels can be derived from numeric loss functions applied along the horizon. Representative loss functions include the simple, average normalized or relative root mean squared error. To draw comparisons with literature results, we suggest the use of Normalized Root Mean Squared Error, NRMSE (7.6) and of Symmetric Mean Absolute Percentage of Error, SMAPE (7.7) [61]. Other less frequent metrics, such as the average minus log predictive density (mLPD), have been shown to hold interesting properties for very specific data contexts.

$$Acc_i(\mathbf{c}_i, \hat{\mathbf{c}}_i) = 1 - NRMSE(\mathbf{c}_i, \hat{\mathbf{c}}_i) = 1 - \frac{\sqrt{\frac{1}{h}\Sigma_{j=1}^{h}(c_i^j - \hat{c}_i^j)^2}}{c_{max} - c_{min}} \in [0,1] \tag{7.6}$$

$$Acc_i(\mathbf{c}_j, \hat{\mathbf{c}}_i) = 1 - SMAPE(\mathbf{c}_i, \hat{\mathbf{c}}_i) = 1 - \frac{1}{h}\Sigma_{j=1}^{h}\frac{|\ c_i^j - \hat{c}_i^j\ |}{(c_i^j + \hat{c}_i^j)/2} \in [0,1] \tag{7.7}$$

7.4.3 Compact Performance Views

An extended confusion matrix to assess multi-period classifiers was proposed in Figure 7.5. The need to collapse some of its axes to facilitate the analysis of results was motivated. In particular, equations (7.4) and (7.5) collapse the temporal axis by computing the average values per metric and label across the horizon of prediction h. However, with this option, we loose the ability to understand which periods are affecting the score. Illustrating, a multi-period learner stable along the horizon of prediction is often preferable over multi-period learner with propensity to classify the first periods without error but whose performance rapidly degrades along the horizon of prediction. With the previous scores, this behavior cannot be assessed. As such, an alternative option is to collapse the labels' axis by defining a predicate with a mapping function T. An illustrative function is one that decides whether an observation is of interest (positive) or not based on the observed values. For example, in healthcare relevant patients can be defined as having at least one hospitalization across the horizon of prediction. In order to avoid the multiplicity of metrics associated with the h periods, the results can be presented under a simple test (based on a fix β-threshold) to evaluate the adequacy of the h predictions for a particular observation, $Acc(c, \hat{c}) \geq \beta$. Understandably, this option comes at a cost of defining a new labeling function T and of optionally working with β-threshold levels. Table 7.1 presents the revised confusion matrix for multi-period classification when two classes are considered. Resulting round accuracy (7.8), sensitivity (7.9) and specificity (7.10) metrics from this setting are also provided.

	Condition Positive (c)	Condition Negative ($\neg c$)
Outcome Positive	tp=$\Sigma_{i=1}^{n}(c{=}T(\mathbf{c}_i)){\wedge}Acc_i(\mathbf{y}_i, \hat{\mathbf{y}}_i){\geq}\beta$	fp=$\Sigma_{i=1}^{n}(c{\neq}T(\mathbf{c}_i)){\wedge}Acc_i(\mathbf{c}_i, \hat{\mathbf{c}}_i){<}\beta$
Outcome Negative	fn=$\Sigma_{i=1}^{n}(c{=}T(\mathbf{c}_i)){\wedge}Acc_i(\mathbf{c}_i, \hat{\mathbf{c}}_i){<}\beta$	tn=$\Sigma_{i=1}^{n}(c{\neq}T(\mathbf{c}_i)){\wedge}Acc_i(\mathbf{c}_i, \hat{\mathbf{c}}_i){\geq}\beta$

Table 7.1: Multi-period confusion matrix

$$RoundAccuracy_c = \frac{tp + tn}{tp + fp + tn + fn} = \frac{1}{n}\sum_{i=1}^{n}(Acc_i(\mathbf{c}_i, \hat{\mathbf{c}}_i) \geq \beta) \tag{7.8}$$

$$Sensitivity_c = \frac{tp}{tp + fn} = \frac{\Sigma_{i=1}^{n}(c = T(\mathbf{c}_i)) \wedge Acc_i(\mathbf{c}_i, \hat{\mathbf{c}}_i) \geq \beta}{\Sigma_{i=1}^{n} c = T(\mathbf{c}_i)} \tag{7.9}$$

$$Specificity_c = \frac{tn}{tn + fp} = \frac{\Sigma_{i=1}^{n}(c \neq T(\mathbf{c}_i)) \wedge Acc_i(\mathbf{c}_i, \hat{\mathbf{c}}_i) \geq \beta}{\Sigma_{i=1}^{n} c \neq T(\mathbf{c}_i)} \tag{7.10}$$

7.4.4 Complementary Evaluation Metrics

Understandably, the proposed loss functions, $Acc_i(\mathbf{c}_i, \hat{\mathbf{c}}_i)$, to evaluate the performance of multi-period classifiers are conservative for the cases where mismatches are caused by temporal lags. Illustrating, although the observed $\mathbf{c}_i$=<0,0,1,1,2,3> and estimated $\hat{\mathbf{c}}_i$=<0,1,1,2,3,4> sequences of labels can be considered dissimilar with regards to the previous scores, in fact, their dissimilarity is simply explained by a soft misalignment associated with a temporal shift. To avoid a significant penalization of the performance of multi-period classifiers when such misalignments occur on the time or value axes, their evaluation can rely on more expressive similarity functions between sequences. Ding et al. [176] and Batista et al. [46] compare the properties of alternative similarity functions when the attribute under classification is ordinal. Dynamic Time Warping (DTW) treats misalignments, which becomes critical when dealing with long horizons. Longest Common Subsequence deals with gap constraints. Pattern-based functions consider shifting and scaling in both the temporal and the amplitude axes.

When the output attribute is nominal, similarity functions proposed to compare biomolecular sequences (based either on their functional or structural similarity) can be applied [436]. These functions are also able to identify temporal shifts as they rely on sequence alignment operators and are, additionally, able to deal with variations on the amplitude axis by detecting character level differences.

On one hand, these similarity functions have the advantage of smoothing error accumulation by allowing temporal misalignments. On the other hand, their use can mask the structural accuracy of multi-period classifiers and lead to more optimistic results.

Additional relevant metrics for multi-period classification include: 1) *error accumulation*, the propagation of past prediction errors, which can be expressed by a bias-variance for squared loss functions; and 2) *smoothness*, the ability of the learner to avoid over-fitting when noise fluctuations are present [129].

7.5 Results and Discussion

The experimental assessment of the proposed contributions is presented in five steps. First, we describe specificities associated with the implementation of the proposed methods and the gathered datasets for the assessment. Second, the performance of multi-period classifiers is evaluated according to the introduced accuracy views and efficiency. Third, the impact from choosing alternative single-label learning settings is briefly assessed. Fourth, we show the relevance of the selected performance view to accurately assess multi-period classifiers. Finally, major implications are synthesized. The experiments were run in an Intel Core i5-2410M 2.30GHz with 6GB of RAM.

Methods. The proposed multi-period methods were codified in Java (JVM v1.6.0-24)[1]. CMPC was tested using two different clustering methods, k-Means [299] and distribution-based Expectation Maximization (EM) [93], in the presence and absence of hierarchical and density-based wrappers provided in Weka [286]. These methods were adopted from Weka extensions BioWeka and TimeSeries [286] in order to rely on distance metrics able to compare the similarity of symbolic time series.

Single-label classifiers prepared to learn from different data structures were selected: 1) a classic classifier, Bayesian networks[2] [517], able to learn either from tabular data or denormalized structured data (by mapping the last m events as features); 2) a sequence classifier, HMMs[3] [478], able to learn either from three-way time series or from structured data when mapped through the use of feature vectors; 3) two advanced classifiers, P2MID

[1] Available in http://web.ist.utl.pt/rmch/software/evoc/

[2] For the used data contexts, this classifier outperformed the performance of kNN lazy learners [11] and C4.5 decision trees [542]. We hypothesize that this is because Bayesian networks are able to implicitly model temporal dependencies from data denormalized from the target structured datasets.

[3] HMMs were chosen due to their maturity, expressive power and inherent simplicity.

pattern-based models and M2ID generative models (see *Chapter VI-4*), able to learn from structured data contexts. Table 7.2 describes the selected parameterizations for the considered classifiers.

Multi-period Classifiers	Parameterized with the same single-label classifiers; Single-output methods: iterative, direct and hybrid (Alg.11) methods with the option to remove labels; CMPC methods with default parameters (Section 3), k-Means with nominal sequence distance and the variances $s \in \{1, 2, 4\}$ and s dynamically parameterized;
Single-label Classifiers	Bayesian networks (optimized parameters through sensitivity analysis) to learn from tabular data; Hidden Markov models (fully-interconnected architecture with 15 states and the Viterbi algorithm) to learn from multivariate time series; Pattern-based classifier P2MID (with default parameters) and generative classifier M2ID (with default extended-left-to-right architecture and the Viterbi algorithm) to learn from structured data;
Data Contexts	Three datasets, each with two-to-three distinct domain mappings (detailed in Table 7.3): tabular, single multivariate time series, and multiple temporal attributes;
Results	10 cross-fold validation; multi-period accuracy (7.2), specificity (7.4) and sensitivity (7.5); Paired two-sample two-tailed t-Student tests with 9 degrees of freedom;

Table 7.2: Selected experimental parameters: algorithmic settings and data settings.

Data Contexts. Three datasets were used to evaluate the proposed multi-period methods: two datasets, msnbc.com[4] and diabetes[5] datasets, from UCI repository [33], and the healthcare heritage database[6] described in *Chapter IV-2*.

The msnbc.com dataset [111] traces sequences of nominal events, where each event corresponds to a user's page request. Events are not recorded at the finest URL level, but using $|C|$=17 categories, such as news, technology, opinion, and health. The goal is to rely on the n=2804 users with more than 50 events and predict the sequence of labels corresponding to the last $h \in \{1..25\}$ page requests from the remaining events. For this purpose, after removing the last h events from data, two single-label learning settings were chosen: HMMs applied over sequences with an average number of $107-h$ events, and classic classifiers applied over a fixed number of features corresponding to the 25 events prior to the horizon of classification. This setting is of interest to study the performance of multi-period methods in the presence of a high number of labels.

The diabetes dataset is a collection of events over a month period for a population of n=64 patients. The events capture the taken insulin doses (regular, isophane or ultra-lente), the glucose measurements and other events of interest, such as hypoglycemic symptoms and exercise activity levels. The selected task is to classify the expected levels for the required daily dose of isophane insuline (NPH) for the last h={1..15} days based on the past events, where the dose levels are Σ={Low(0-9),Normal(10-17),High(18-25),VeryHigh(25+)}. Since several insulin doses are daily taken per patient, we performed the daily sum of insulin doses per type (regular, NPH and ultra-lente) and the daily average of glucose measurements, which results in 4 daily events. After removing the last h days from data, two single-label learning settings were chosen: HMMs applied over the multivariate sequences of daily events (average number of $397-4h$ events per sequence), and classic classifiers applied over 65 features corresponding to daily events (4×15 features) and additional events of interest (5 features) from the 15 days prior to the time horizon of classification.

Finally, healthcare heritage prize database is a large-scale database that integrates healthcare claims across hospitals, pharmacies and laboratories for n>150000 patients. The original relational scheme was mapped in a dimensional scheme, where each patient has a registry of the claims and of the monthly number of laboratory tests and taken drugs. We selected the planning task of classifying the level of prescription needs ($|C|$=5) for sequent periods, considered to be critical for healthcare prevention and drug management. For this task we selected month and quarter temporal granularities, and varied $h \in \{1..12\}$, removing the events that occurred in this time horizon. We rely on three distinct data mappings to study the impact of alternative learning settings in the multi-period classification task. Their details are provided in Table 7.3.

[4] http://archive.ics.uci.edu/ml/datasets/MSNBC.com+Anonymous+Web+Data
[5] http://archive.ics.uci.edu/ml/datasets/Diabetes
[6] http://www.heritagehealthprize.com/c/hhp/data (under a granted permission)

Setting	Statistics
Denormalized data (tabular representation)	Single-value attributes were maintained. The last five (multivariate) claims and all monthly occurrences were denormalized as single-value attributes. Missing labels were used to fill absent claims (patients with low clinical activity). $\mathcal{A}$ has 128 attributes for the month granularity (108 events and 20 static) and 88 attributes for the quarter granularity.
Multivariate time sequence representation	Monthly mode and counting aggregations were applied for 'procedure' and 'diagnosis' fields from claims to compose a time series with p=4 order, and combined with monthly lab tests. Resulting time series per patient has 24-to-30 time points and p=5 order.
Time-enriched itemset sequence	Details described in [302]. For the month granularity, each patient has an average number of 4 items per itemset (σ=2) and 36 itemsets per sequence. For the quarter granularity, patients have an average number of 12 items per itemset (σ=3) and 12 itemsets.

Table 7.3: Statistics for the healthcare heritage dataset after pre-processing.

7.5.1 Comparing Multi-period Classifiers

Structural Performance. Table 7.4 provides an overview of the performance of the proposed CMPC method over the introduced data settings against two baseline alternatives: iterative and direct single-output strategies adapted from long-term prediction. We selected an horizon with h=12 periods and the CMPC-MISMO method with variance s=4 for this analysis. The p-values computed from testing the superiority of CMPC over the baseline single-output methods (in terms of accuracy) are provided for the different data settings. This analysis is complemented by the results provided in Table 7.5 that gather the observed accuracy levels of single-output methods and multiple-output methods for varying parameters. Additionally, we also disclose the observed differences in sensitivity, $\triangle sensitivity$, against direct single-output methods.

Data (h=12)	Data domain (single-label learner)	CMPC accuracy	Nr. of clusters	direct accuracy	p-value (CMPC *vs* direct)	iterative accuracy	p-value (CMPC *vs* iterative)
msnbc.com	tabular (Naive Bayes)	0.64±0.02	38	0.61±0.02	0.01	0.61±0.02	8E-3
	m=1 time series (HMMs)	0.63±0.02		0.59±0.02	7E-3	0.60±0.02	0.02
diabetes	tabular (Naive Bayes)	0.55±0.06	14	0.50±0.04	0.08	0.49±0.04	0.07
	m=4 time series (HMMs)	0.59±0.05		0.52±0.03	0.01	0.52±0.04	0.02
healthcare	tabular (Naive Bayes)	0.88±0.01	11	0.85±0.01	3E-3	0.85±0.01	1E-3
	m=5 time series (HMMs)	0.86±0.02		0.83±0.02	0.05	0.83±0.02	0.04
	multi-attribute (M2ID)	0.92±0.02		0.87±0.02	2E-3	0.87±0.02	5E-3

Table 7.4: Performance of CMPC-MISMO (s=4, iterative behavior and default parameters) against direct and iterative single-output methods for different datasets and single-label learners using t-tests.

		msnbc.com (times series)		diabetes (times series)		healthcare (multi-attribute)	
method	sVariance (h=12)	accuracy	△sensitivity (*vs* direct)	accuracy	△sensitivity (*vs* direct)	accuracy	△sensitivity (*vs* direct)
random	1	0.06	–	0.25	–	0.25	–
combinatorial	12 (MIMO)	0.32±0.12	0.33	0.31±0.13	0.20	0.47±0.10	0.24
combinatorial	4 (MISMO)	0.49±0.08	0.27	0.44±0.07	0.13	0.79±0.06	0.17
direct	1	0.59±0.02	–	0.52±0.03	–	0.87±0.02	–
iterative	1	0.60±0.02	0.17	0.52±0.04	0.04	0.87±0.02	0.05
hybrid	1	0.61±0.02	0.13	0.53±0.04	0.03	0.87±0.01	0.05
CMPC	12 (MIMO)	0.62±0.02	0.27	0.64±0.02	0.07	0.91±0.02	0.16
CMPC	4 (MISMO)	0.63±0.02	0.24	0.59±0.05	0.07	0.92±0.02	0.14
CMPC	dynamic	0.63±0.02	0.25	0.59±0.05	0.64	0.92±0.02	0.14

Table 7.5: Accuracy and differences on sensitivity (averaging absolute differences per label against direct methods) for single-output and multiple-output methods (combinatorial and CMPC with varying variances) for data settings with h=12.

Before entering the discussion, two notes may support the analysis of these results. First, by comparing the performance of multi-period methods with the converging performance of a random classifier ($|C|^{-1}$), we observe that the complexity of the learning tasks varies for each dataset. For instance, classifying insulin dose levels is complex as it largely depends on unrecorded contextual factors, while anticipating the sequence of visited page's categories from the msnbc.com dataset is more adequately described by previous events. Second, although the

values of accuracy are similar for some pairs of multi-periods methods (such as direct and iterative methods or CMPC's MIMO and MISMO methods), their behavior can significantly differ. This observation is shown by the observed differences in the computed sensitivity among classes.

In this context, four major observations can be retrieved from these tables. First, CMPC methods have significant accuracy improvements for the majority of the considered datasets and mappings. CMPC methods support the learning of non-trivial sequences of labels as it is demonstrated by the number of parameterized clusters against the number of labels ($\gg |C|$). We hypothesize from the empirical results that CMPC methods can minimize the error accumulation of iterative single-output methods while avoid the restrictive independence among periods of direct single-output methods. Although the absolute gains in accuracy are moderate for the msnbc.com and healthcare datasets, the observed differences are statistically significant as the performance across folds is considerably stable. The significance of the differences between CMPC and its peer methods increases for larger horizons h[7]. Second, we observed that CMPC-MISMO variants slightly improve accuracy. Let us consider the diabetes dataset. The prototype combinations of labels (sequential behavior) associated with these h/s periods (given by cluster's centroids) promote a more focused learning task, while are able to capture periodicities on the levels of taken insulin doses. Now consider the msnbc.com dataset. The previous observation is no longer true due to the non-ordinal nature of labels and high cardinality $|C|$. In this data context, the number of estimated clusters tends to be large, >30, and each one of these clusters are learned from a similar number of observed observations (near 100). An interesting property is that multiple page requests under the same category lead to sequence of events with the same label prior to a page request on a different category, e.g. $\mathbf{c}=\{3,3,3,3,4,4,5,5,5,5\}$ for $h=10$. In this setting, single-output methods tend to output the same label across all the periods under classification. Contrasting, since CMPC is learned in a setting where some clusters contain variations on the visited pages (as the number of clusters is higher than $|C|$), it is more able to learn sequences with multiple labels. Third, the differences in sensitivity against direct methods underline the different behavior of these strategies. In particular, CMPC methods are able to deal with the imbalance (associated with the low representativity of some labels in these datasets) and thus minimize the risks of overfitting since the learned clusters can group infrequent sequences with closer sequential behavior. Finally, the performance of multi-period classifiers significantly varies with the chosen single-label learners since they rely on distinct data mappings that lead to different learning complexities.

Comparing Accuracy. Figure 7.6 provides a closer look on how the accuracy of multi-period classifiers varies with the horizon of prediction for a classic single-label classification setting (Bayesian networks over denormalized data). Combinatorial multiple-output strategies adapted from multi-label classification are not visually represented since their accuracy rapidly degrades with h for the diabetes dataset (due to the low number of observations) and msnbc.com dataset (due to the high number of labels). Still, their accuracy has significant improvements against the converging accuracy of a random classifier. When comparing the alternative methods, three major observations can be retrieved. First, for short horizons of prediction ($h\leq 5$), the performance of CMPC is better than combinatorial methods, while competitive with single-output peers (direct and iterative methods). A closer analysis of the behavior of CMPC for a small number of periods reveals that the number of dynamically selected clusters tends to be equal or slightly greater than the number of classes ($|C|$), thus, mimicking the behavior of single-output methods. Second, the CMPC methods perform better than single-output methods for larger horizons. In particular, the accuracy of the MISMO variant is statistically superior over single-output methods for the diabetes dataset when $h\geq 15$ (assuming $s=5$), for the msnbc.com dataset when $h\geq 12$ (assuming $s=4$) and for the healthcare dataset when $h\geq 8$ (assuming $s=4$). Although the absolute differences in accuracy appear to be subtler for the two latter datasets, they stand on lower levels of variance across folds. Third, the constrained inclusion of past predictions through the proposed hybrid approach is as competitive as the best single-output strategy and, therefore, is the preferable single-output option.

[7] Complete list of results available in http://web.ist.utl.pt/rmch/software/evoc/

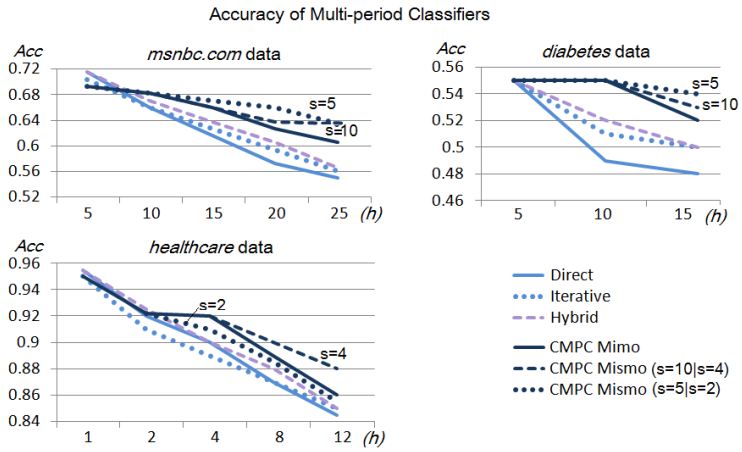


Figure 7.6: Comparing the accuracy ((VI-7.2) parameterized with (VI-7.3)) of direct, iterative, hybrid and CMPC approaches for the diabetes, msnbc.com and healthcare datasets using Bayesian networks after data denormalization. Combinatorial strategies were excluded as the low levels of performance impact interpretability of the chart.

Comparing Sensitivity. An analysis of the levels of sensitivity (7.4) across multi-period methods maintaining the same experimental setting is provided in Figure 7.7. The levels of sensitivity per label vary substantially due to the differences on label-conditional learning complexity. Illustrating, the labels related to medium-to-high levels of drug prescription (healthcare dataset) and of insulin doses (diabetes dataset) are related with patients with less stable behavior. A major observation from this analysis is that multiple-output and single-output methods have distinct sensitivities per label, which underlies the impact of this choice since outputs can significantly differ. A relevant fact is that the CMPC method slightly balances the sensitivity levels across labels. This is because frequent sequences of labels are separated in nearly equiprobable spaces through the use of clusters. Consequently, CMPC classification models are more able to classify new observations with less probable labels (e.g. medium levels of drug prescription) than single-output peers. This an important achievement since less probable labels are commonly labels of interest (positive labels).

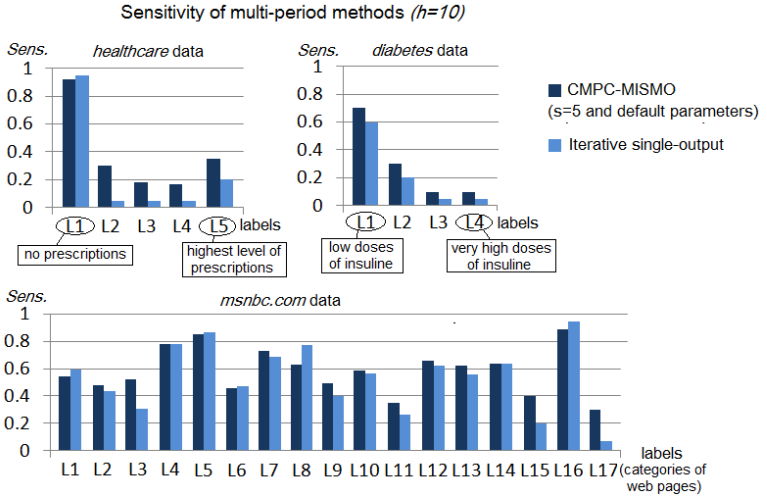


Figure 7.7: Comparing the sensitivity (according to (VI-7.4)) of CMPC MISMO and iterative single-output methods (with Bayesian networks) for the three datasets with h=10.

Impact of Algorithmic Choices. A non-exhaustive view on how CMPC varies with algorithmic and distance choices for the healthcare dataset is provided in Table 7.6. For a simplified analysis of the confusion matrix, we grouped labels according two major levels of prescriptions: low level and moderate-to-high levels. The k-Means approach holds slightly better levels of accuracy than its peers. Density-based wrappers or the EM algorithm [93] are able to increase the number of true positives and decrease the number of false positives. However, these options are easier prone to a slight increase in the number of false negatives. The use of similarity metrics between time series slightly increases performance against traditional distance metrics that cannot properly score misaligned labels.

Parameter (tabular healthcare dataset; h=12; $\Sigma_{positive}$={2,3,4} $\Sigma_{negative}$={0,1})	accuracy	sensitivity	specificity
CMPC with default parameters	0.88±0.01	0.53±0.03	0.94±0.01
CMPC with density wrapper	0.87±0.02	0.54±0.03	0.92±0.02
CMPC with hierarchical wrapper	0.85±0.02	0.51±0.03	0.91±0.02
CMPC with classic Euclidean distance (instead of sequences' similarity)	0.86±0.01	0.51±0.03	0.92±0.01

Table 7.6: Impact of the different algorithmic and distance choices on the behavior of the multi-period CMPC-MISMO method with s=4 over the denormalized healthcare dataset.

Comparing Efficiency. Figure 7.8 compares the time efficiency (clock units) of the different multi-period classifiers for the healthcare dataset. The proposed CMPC methods have a distinctive better performance. Understandably, the efficiency increases with s variance ($\propto s$) for the MISMO models due to a decrease in the number of iterations. Contrasting, hybrid approaches are the less efficient alternatives since they have to compute both direct- and iterative-based classifiers for each period. Single-output methods have a degrading complexity performance with an increased horizon as the number of algorithmic iterations linearly increases. For very large horizons ($h \gg 12$), the use of multiple-output methods with lengthy temporal partitions become critical and decisive to guarantee the scalability of multi-period classifiers.

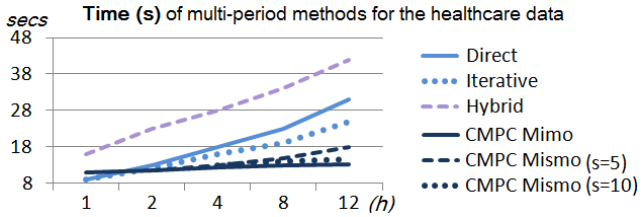


Figure 7.8: Comparing the time efficiency of multi-period methods (with Bayesian networks) over the healthcare data.

Impact of Single-label Learning. Figure 7.9 presents the accuracy of the CMPC method (s=2) under alternative learning settings for the diabetes and healthcare datasets. Single-label classifiers that model temporal and cross-attribute dependencies from the data domain $\mathcal{X}$ promote significant improvements on the multi-period classification task. Understandably, these improvements are associated with the ability of single-label classifiers to relate events in the data domain and weight their relevance according to their time of occurrence. A complementary and more detailed discussion on the performance of alternative single-label learners in data domains with multiple temporal attributes can be found in [302].

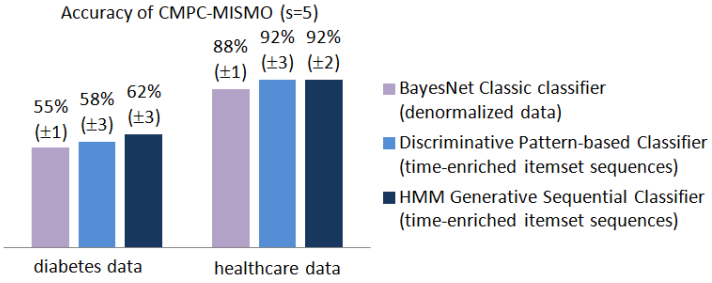
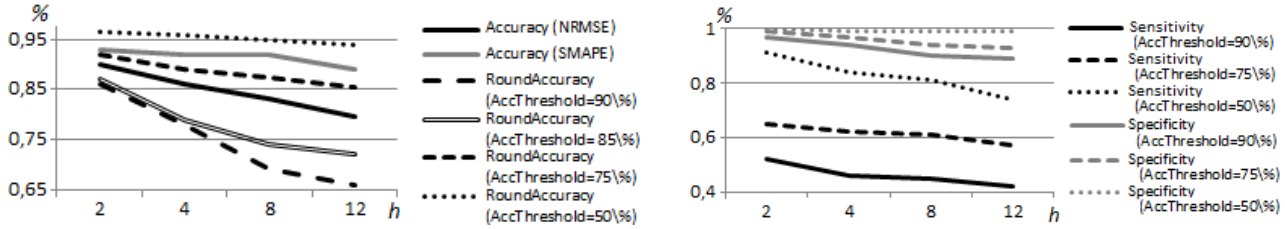


Figure 7.9: Comparing the impact of distinct single-label classifiers for the heritage dataset with h=10 using CMPC MISMO with s=5. Selected settings: classic classifier (Bayesian networks) from denormalized data, discriminative pattern-based classifier (P2MID) and generative classifier (M2ID) from time-enriched itemset sequences.

Understanding Alternative Evaluation Metrics. Figure 7.10 compares alternative performance views – the round accuracy (7.8), sensitivity (7.9) and specificity (7.10) – parameterized with different loss functions and β-thresholds for an average multi-period classifier[8]. Since the attribute under prediction, number of prescriptions, is ordinal, we considered an average mean accuracy (7.2) using the normalized root mean squared error (NRMSE) (7.6) and the symmetric mean absolute percentage of error (SMAPE) (7.7) loss functions. We varied the accuracy threshold β

[8]Results are the average of direct, iterative, hybrid and CMPC (with sampling variances s={2, 4})

from 50% to 90%. Additionally, we considered a relevance criterion based on the number of drugs prescribed per patient, $T(\mathbf{c}) = (\Sigma_{i=1}^{h} c_i) > h$. Exemplifying, given an observation with the following observed and predicted values $\{c=(2,4,6), \hat{c}=(3,4,1)\}$ with $|C|=\{0,1,..,6,7+\}$, this observation is relevant since $T(\mathbf{c})=12 > 3$ and the prediction is non-accurate since $Acc_{NRMSE}(y)=57<85\%$.



(a) Alternative accuracy views: round (7.8) and average (7.2).

(b) Compact sensitivity (7.9) and specificity (7.10) views.

Figure 7.10: Comparing alternative performance views of an average multi-period classifier for classifying the levels of drug prescription using the heritage dataset.

Three major observations are retrieved. First, when comparing loss functions, NRMSE alleviates small differences between the predicted and observed values and penalizes larger differences. SMAPE curve has higher scores since large differences are easily masked when the distances are normalized. Second, round accuracy (7.8) values are worse than simple accuracy (7.2) for thresholds above 75% due to a considerable number of observations with an accuracy slightly below the fixed β-thresholds. An analysis of round accuracy curves for a fixed horizon length with varying β-thresholds can be used to disclose the ranges where a significant number of true decisions becomes false. Such analysis is critical for real-life planning tasks where decisions, as the need for a specific treatment, need to be made with certain confidence. Finally, sensitivity is substantially low for accuracy thresholds above 90%, although it increases to more acceptable levels under more relaxed accuracy thresholds. This seems to be partially explained by the natural variability of predicting drug prescriptions based on clinical claims and patient profile.

7.5.2 Discussion

A set of implications can be synthesized from the presented set of observations. First, CMPC methods are the natural option for medium-to-large horizons of prediction, with the s-variance being dependent on the extent of local sequential dependencies and approximately determined by the number of periods and labels. Illustrating, given the healthcare dataset with $|C|=5$, $s=$h is the best option for $h \in \{2,3,4,5\}$, while $s=h/2$ is the best option for $h \in \{6,8,10\}$ and $s=$h$/4$ for $h \in \{12\}$. On one hand, CMPC can suffer from smoothing for datasets with a low number of observations since the dynamically selected number of clusters that guarantees an adequate number of observations per cluster ($\approx \frac{m}{5}$) is not sufficient to capture the diversity of sequential behavior. An analysis of the clusters learned over the diabetes dataset for a large number of periods revealed that many sequences of labels of potential interest were neglected. Still, the focus on a specific subset of sequences of interest for classification can benefit the learning (see Figure 7.6) since it alleviates the problem of having a prohibitive number of sequences associated with combinatorial methods. On the other hand, for datasets with a medium-to-high number of observations, CMPC provides an adequate learning of non-trivial sequential behavior without the risk of overfitting. This behavior is supported by the analysis of the output sequences (with frequent combinations of different labels) and by the improvements for the sensitivity levels for less frequent labels.

The proposed CMPC methods have the additional advantage of being able to extend the behavior of any single-label classifier without the need to adapt the core learning task. The observed levels of accuracy from Tables 7.4 and Figure 7.9 support the hypothesis that defining new single-label classifiers able to learn from temporal domains, such as the domains characterized by multiple time sequences and static attributes, is critical to obtain distinctive levels of performance for the multi-period classification task.

The use of complementary performance views is important to further tune these methods or assess the impact of novel methods. For some data settings, the classification can be prone to temporal misalignments (e.g. need for

hospital intervention depends on the individual pace of disease progression), which (7.2) does not account for. In this context, the use of advanced similarity metrics for both nominal and ordinal labels (see *Section 7.4.4*) should be considered for more fair and less conservative assessments.

7.6 Summary of Contributions and Implications

This work introduces the emerging task of multi-period classification and proposes new methods able to surpass the inherent limitations associated with methods adapted from long-term prediction and multi-label classification. Related research streams were surveyed according to the two major requirements of the multi-period classification task: 1) independence of the single-label learning setting (enabling the use of previous proposed principles to guarantee the accuracy and significance of classification decisions), and 2) ability to model the conditional dependencies between the periods under classification.

A novel multiple-output approach relying on clustering methods to capture prototype sequences of labels, CMPC, was proposed. This approach is able to preserve local stochastic dependencies without the need to adapt the underlying single-label classifier. Additionally, the use of cluster's centroid for space reduction and recovery is an effective strategy to avoid the problems of combinatorial strategies and to guarantee an optimal learning that is dynamically parameterized based on the number of periods, number of labels, number of observations and combinations of labels. This parameterizable behavior is critical to jointly minimize the over/underfitting propensity of the learning function towards data with a limited number of observations.

The conducted experiments show the superior performance of the proposed CMPC variants for distinct real data in terms of efficiency and accuracy. We also show that cluster-centric behavior of the CMPC method is able to guide the learning towards non-trivial sequences of labels without overfitting risks. This leads to improvements in the sensitivity of less frequent labels.

Future work. We identify five major directions for future research. First, we expect to see the proposed (and upcoming) multi-period classifiers applied to answer a wide-set of increasingly prominent real-world problems (listed in *Section 7.1.2*). Furthermore, since the proposed formulation of the multi-period task was shown to comply with the prediction of upcoming h events and the classification of non-uniform and non-convex periods, we also aim to demonstrate its adequacy towards these alternative ends.

A second possible direction is to further explore principles to adequately classify high-dimensional sequences of labels (e.g. memory sampling by selectively forgetting some of the estimated periods). In this context, we identify two important topics of research. First, being able to frame confidence of the classified periods based on the longevity of the period under prediction. Second, understanding when and how local-stationarity can be assumed to guide the selection of the s-variance parameter required by multiple-output methods.

A third direction of interest is to rely on updatable multi-period classifiers to anticipate critical events based on continuously incoming data. To guarantee that single-output classifiers given by direct and iterative methods are updatable, we need to simply guarantee that the underlying single-label classification model is updatable. To guarantee that CMPC is updatable, we need to guarantee that both the clustering method and the underlying single-label classification model are updatable.

Fourth, we expect to extend the studied multi-period principles for learning settings where the attribute under prediction is numeric. Although CMPC can be easily adapted for clustering numeric time series, CMPC is not able to rely on classic regression models since clusters have not ordinal relations.

Finally, we expect to study the impact of using more expedite similarity functions to: 1) gain further insights of the true performance of multi-period classifiers by considering metrics able to smooth (possibly) ordinal and temporal misalignments; and 2) affect the behavior of multi-period classifiers by revising the underlying metrics considered by the clustering methods.