UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Learning from High-Dimensional Data using Local Descriptive Models

THESIS SUMMARY

Rui Miguel Carrasqueiro Henriques

Supervisor: Doctor Sara Alexandre Cordeiro Madeira

Thesis approved in public session to obtain the PhD Degree in
Information Systems and Computer Engineering

Jury final classification: **Pass with Distinction and Honor**

Jury

Chairperson:	Chairman of the Scientific Board
Members of the Committee:	Doctor Mário Alexandre Teles de Figueiredo
	Doctor Joaquín Dopazo Blásquez
	Doctor Miguel Francisco de Almeida Pereira da Rocha
	Doctor Francisco João Duarte Cordeiro Correia dos Santos
	Doctor Sara Alexandre Cordeiro Madeira

2016

Abstract

Models learned from high-dimensional data, where the high number of features usually exceeds the number of observations, have higher propensity to either overfit or underfit data. In this context, it is thus important to focus the learning on regions of interest, such as subsets of features, guaranteeing that these regions are both informative and statistically significant. Although a composition of relevant regions can be learned under specific assumptions to offer these guarantees, the state-of-the-art learning methods place restrictive constraints on the allowed structure, coherency and quality of regions. This has prevented the understanding of how the properties of the selected regions affect the performance of descriptive and classification methods in both tabular and structured data contexts.

In this work, we propose robust, flexible and statistically significant local descriptive models and study their relevance to improve (associative) classification in high-dimensional data contexts. This task is tackled in three major steps. *First*, we propose new local descriptive models from tabular and structured data with robustness and flexibility guarantees. In the presence of matrices and network data, the focus is placed on learning biclustering models able to tackle existing challenges: learn from regions with flexible coherency (additive, symmetric, plaid and order-preserving models); guarantee scalable searches; robustness to varying forms and degree of noise; model regions from sparse data; and effectively incorporate background knowledge. In the presence of structured data, possibly given by multivariate time series or multi-sets of events, the focus is placed on new deterministic and generative methods to learn local descriptive models given by cascades of modules or arrangements of informative events. *Second*, we propose principles to both assess and guarantee the statistical significance of these descriptive models. *Third*, the previous contributions are extended towards labeled data contexts, and new training and testing functions are proposed to learn associative classification models. In this context, we assess the impact of varying structures, coherencies and quality of local descriptive models on the performance of classifiers, and combine statistical significance and accuracy views to study and revise their behavior. Finally, we extend these contributions for data with structured classes to adequately answer predictive tasks.

The proposed contributions were applied to tackle a wide-set of real-world tasks in biomedical and social domains, including the learning of descriptive and predictive models from gene expression data, repositories of health records, clinical data, collaborative filtering data, and (biological and social) networks.

Keywords:
High-Dimensional Data
Structured Data
Biclustering
Local Descriptive Models
Associative Classification
Biomedical Data Analysis
Statistical Significance
Multivariate Time Series
Multi-Sets of Events
Sparse Data

I. Foundations

Learning from high-dimensional data, where the high number of features can exceed the number of observations, is challenged by an inherent complexity and generalization difficulty. In these data contexts, these challenges can be minimized by focus the learning on specific regions of interest (such as subsets of features) [17]. However, the lack of flexibility on how the existing learning methods select such regions is associated with three major problems: *1)* the inclusion of non-relevant regions (promoting overfitting), *2)* the exclusion of relevant regions (promoting underfitting), and *3)* the modeling of apparently relevant regions, yet not statistically significant [17, 3, 22]. As illustrated in Figure 1, learning from high-dimensional data is a challenging task since not all regions are equally informative and, even when informative, regions may not be statistically significant. Furthermore, they can be significantly informative yet non-significantly discriminative.

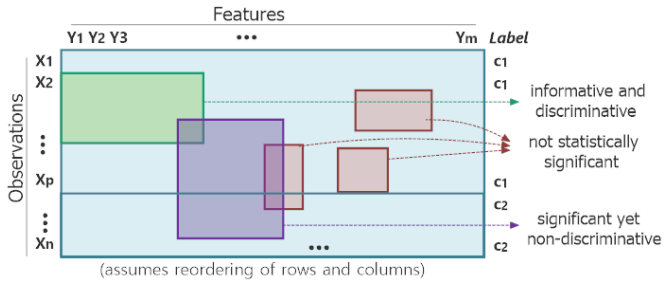

Figure 1: Learning from high-dimensional data: relevance of selecting coherent, discriminative, significant regions.

This work aims to tackle these challenges by learning flexible, robust and statistically significant descriptive models and associative classification models from relevant regions of a high-dimensional data space. This learning task is addressed for tabular and structured data. In this context, our goal is to systematically study how does the performance of descriptive and classification models vary with the homogeneity, discriminative power and statistical significance of the selected regions. The underlying *hypothesis* is that the adequate selection and composition of regions improves the performance guarantees of local descriptive models and (associative) classifiers learned from high-dimensional data. As a result, this understanding opens a window of opportunity: new principles can be inferred to revise the behavior of existing learning methods.

The importance of this thesis is driven by two major observations. First, the need to validate the increasing number of scientific statements from the analysis of high-dimensional data without proper statistical assessments [17]. This is particularly critical across biomedical domains, due to the severity of implications that some statements might have on human health and upcoming research. Second, the need to face the increasing dimensionality of the available data, without being susceptible to the problems of feature selection and peer procedures for dimensionality reduction. The focus on a small subset of features is commonly associated with the exclusion of relevant regions and inclusion of non-relevant elements, contributing to the over/underfitting risk.

The in-depth study of how to optimize the performance of the target learning methods, while guaranteeing their statistical significance, is thus critical to answer a wide-set of real-world learning problems. In this work, we address the tasks of learning from: *1)* tabular data from biomedical domains with a high number of molecular units or clinical features per sample or patient, and social domains with a high number of rated items, traits or behavioral features per subject; *2)* weighted graphs given by large-scale biological and social networks; *3)* sequential data associated with (multivariate) time series with a high number of time points and/or (sliding) features; and *4)*

structured data mapped from high-dimensional multi-sets of events, such as repositories of health records, trading decisions, (e-)commerce operations and browsing events.

Below, we explore the current limitations and opportunities of learning from high-dimensional data; structure the problem space according to its requirements; provide a high-level view on the contributions of the thesis; and, finally, provide a roadmap for an easy exploration of the contents in the dissertation.

Problem Motivation

Figure 2 lists the major challenges and commonly applied principles to learn from high-dimensional data.

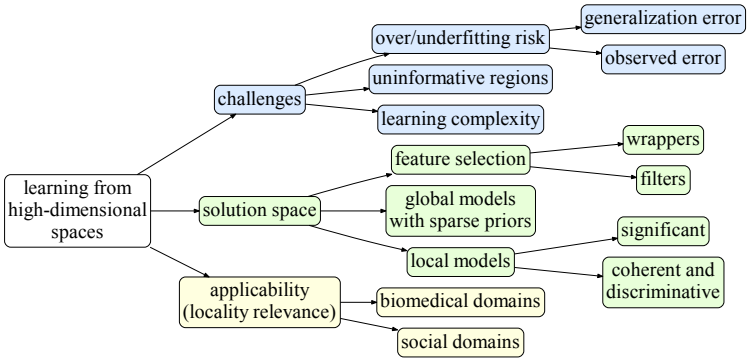

Figure 2: Motivating the learning in high-dimensional spaces: open challenges, contributions and applications.

A well-known challenge of learning from high-dimensional data is associated with the propensity of the resulting models to either overfit or underfit the observed data [30, 22]. Minimizing this risk requires essentially an optimal trade-off between the *observed error* – error estimated from assessing the learned model on the observed data –, and the *generalization error*, often given by the mean and variability of the error estimates collected from assessing the model on new sets of observations [8]. In this context, adjusting the complexity (also referred capacity) of the learning function is necessary to achieve good generalization. Complex functions guarantee a low observed error but often perform poorly on unseen data (overfitting propensity), while overly simple functions may not be able to model relevant data regularities (underfitting propensity).

However, in data contexts where the number of features exceeds the number of observations, the complexity term cannot be explored since models may not be able to generalize. This property is often referred as perfect overfitting towards the observed data [30]. To illustrate this problem, let us consider the following simplistic global model: a linear hyperplane $M(\mathbf{x})$ in $\mathbb{R}^m$ defined by a vector $\mathbf{w} \in \mathbb{R}^m$ and point b to either separate two classes, $sign(\mathbf{w} \cdot \mathbf{x} + b)$, predict a real-valued outcome, $\mathbf{w} \cdot \mathbf{x} + b$, or describe the input observations, $\mathbf{X} \sim \mathbf{w} \cdot \mathbf{x} + b$. As illustrated in Figure 3, a linear hyperplane in $\mathbb{R}^m$ can perfectly model up to $m + 1$ observations, either as a global classifier $\mathcal{X} \rightarrow \{\pm 1\}$, as a regression model $\mathcal{X} \rightarrow \mathbb{R}$ or as a global descriptive model of $\mathbf{X}$. Although the assessment of these models using the same observations is associated with a zero observed error, in the presence of new observations, the generalization error can be significantly high due to the risk of perfect overfitting towards the training data.

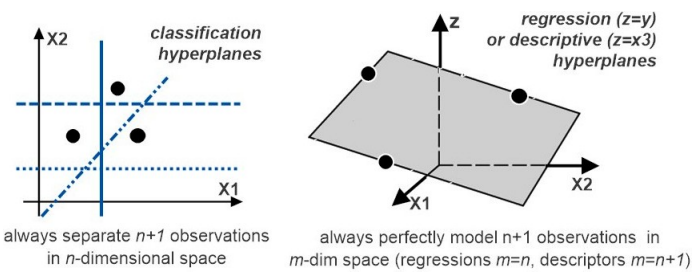

Figure 3: Linear hyperplanes cannot generalize when the number of features (data dimensionality) is larger than the number of observations (data size), $m \geq n + 1$.

As a result of these observations, learning from specific regions of interest from high-dimensional data has been presented as an option to avoid perfect overfitting. However, assessing the statistical impact of selecting regions is critical since small regions are highly prone to be relevant by chance [17].

In tabular data contexts, regions given by a small number of features and/or observations can be highly coherent (descriptive tasks) or highly discriminative (classification tasks), yet their probability of occurrence might not be statistical significant. Illustrating, consider a real-valued matrix defining a (n=50,m=10000)-space with an Uniform distribution of values per feature $\mathbf{y}_j$~U(-1,1) and two balanced classes. Consider a region given by a subset of the original features defining a (50,5)-space. The combination of values for the selected subset of 5 features (assuming an entropy ratio above 90%) is highly likely to occur by chance and therefore this region is not statistically significant. Alternatively, let us neglect the labels and consider a (n=20,m=10)-space. The probability that this region to have constant values across observations is 44% assuming a simplistic binomial calculus, and thus not statistically significant. These problems are similarly observed in structured data contexts, where a region is additionally associated with a subset of time points, item occurrences or events.

Although the selection of small regions is highly prone to be either informative or discriminative by chance, many classifiers: 1) rely on feature selection to deal with high-dimensionality, or 2) infer decisions from regions given by (possibly small) subsets of features. Illustrative classifiers with propensity towards this behavior are decision trees. Decision trees typically select a small subset of features, whose combination of values is able to discriminate a specific class (even if discrimination ability is observed by chance). As a result, decision trees and peer classifiers show a high generalization error.

Understandably, the selection of non-significant regions is associated with the risk of underfitting the observed data. This is the tackled problem in this work since this risk is not structural, meaning that it can be minimized. For this aim, the impact of mapping an original data space into a set of regions needs to be addressed.

In addition to this problem, the selection of uninformative regions increases the learning complexity and can introduce unnecessary biases on the learned models.

Problems of Dimensionality Reduction. Let us further explore the facets of this problem. Three major learning options have been considered for the learning of descriptive and classification models from high-dimensional data.

First, *feature selection* methods have been applied as a filter or a wrapper. Filters select subsets of features as an independent preprocessing stage according to some measure of feature relevance, which often neglects the statistical significance of the selected spaces [32]. Wrappers can be alternatively applied to minimize this problem since they can estimate the generalization error of the model by evaluating the learned model against multiple subsets of features [12]. However, minimizing the generalization error does not guarantee that the selected subsets of features are statistically significant. Additionally, wrappers degrade the learning efficiency and are dependent on the chosen model, that is, there are no guarantees that a subset of features chosen for one model is adequate for other models.

In real-valued data contexts, an alternative simplistic way of reducing the dimensionality of a given dataset is to use a mapping function, also referred as a projection or hyper-dimensional transformation, from the observed data space into a new data space with lower dimensionality $\phi : \mathbb{R}^m \rightarrow \mathbb{R}^d$ where d<m. Contrasting with feature selection, projections can affect the value distributions of features, thus often facilitating the subsequent learning task. However, even in the presence of complex mapping functions, these procedures are not able to flexibly select an arbitrary number of regions from the input data space [17].

Second, and complementarily to feature selection, global models can be learned using *sparse kernels*. A sparse kernel is a parametric learning function that it is able to guarantee a focus on relevant regions by placing assumptions to collapse or disregard parameters associated with uninformative regions, thus minimizing the learning complexity and fostering the model's generalization [5]. Sparse kernels are often associated with (but not limited to) the learning of probabilistic models [11, 17]. In these contexts, irrelevant and redundant parameters rapidly

converge to zero. For this end, specific a priori knowledge regarding the data regularities, referred as prior, have been used to promote sparsity of both unstructured and structured models for both descriptive and classification tasks. Although the covered sparse kernels offer the possibility to discard non-informative data and to balance the over/underfitting by controlling the number of iterations associated with the learning of a parametric model, they show some inherent challenges. Since sparsity is determined by the model's parameters, it is not expressive enough to guarantee a flexible selection of regions of interest due to two major challenges. First, although recent contributions can be used to avoid the need to specify or estimate the degree of sparseness of the models [13], there is still a high complexity associated with the definition of sparse priors. Second, sparsity is primarily used to either discard non-relevant features and/or specific ranges of values per feature, thus preventing the flexible selection of subsets of both observations and features/events/time points.

Third, some learning functions infer descriptions and decisions from sets of regions of the original data space. These functions are associated with local descriptive models, such as biclustering models, and local decision models, such as associative classification models. The problem of how to guide the learning of these models to adequately select and compose regions of interest is the central task of this work. Naturally, the implications of this study can be further used to assess and extend methods for dimensionality reduction (including but not limited to feature selection) as well as to affect the learning of global models.

Relevance of Local Models: Applications. To further motivate the relevance of learning local descriptive models, Table 1 provides a set of biomedical and social data domains characterized by the presence of meaningful local regions. These data domains are characterized by a high-dimensionality associated with a high number of genes per sample, health-records per patient, molecules per biological network, time points per physiological signal, browsing actions per user, trading decisions per business, or interactions per user in social contexts.

	Data	*Illustrative regions with relevance for learning tasks*
Biomedical	physiological [4, 9]	Sets of (sliding) features and signal partitions with coherent values across case or stimuli-elicited responses.
	clinical [15, 18]	Groups of patients with correlated clinical features or health records (shared treatments, diagnoses, prescriptions).
	structural variations [10]	Correlated groups of mutations and copy number variations.
	biological networks [21]	Modules of genes, proteins or metabolites with meaningful interaction (from matrices with pairwise connections).
	gene expression [16, 26]	Groups of genes involved in functional processes and pathways only active under certain conditions.
	genome-wide [31, 29]	Conserved functional subsequences (sequence alignments), factor binding sites and insertion mutagenesis.
	other	Local regularities in translational [7], chemical [25] and nutritional data [24].
Social	social networks [14]	Groups of individuals with correlated activity and intercommunication; groups of contents based on accessors.
	text mining [2, 26]	Content-related documents and web pages (from matrices weighting categories/words across text segments).
	(e-)commerce [1]	Hidden browsing patterns containing relationships between sets of (web) users, (web) pages and operations.
	financial trading [23]	Indicators producing similar profitability for specific trading points (buy, hold and sell signals) in the stock market.
	collaborative filtering [6]	Groups of users who share preferences and behaviorial patterns for a subset of available actions.

Table 1: Disclosing the meaning of regions across (high-dimensional) biomedical and social data contexts.

Thesis Requirements

The underlying **hypothesis** is that *learning from relevant regions of high-dimensional data improves the performance guarantees of local descriptive models and (associative) classification models*. Naturally, testing this hypothesis leads us into the *how*. First, how does performance vary with the properties of the selected regions? Second, how can this understanding be used to improve the learning of descriptive and classification models? Figure 4 lists the key requirements and premises to validate the target hypothesis.

First, in order to validate the proposed hypothesis, we decompose its assertion according to an incremental set of five major requirements. These requirements define the problem space.

R1 Robust assessment of descriptive and classification models learned from high-dimensional data.

By satisfying the first requirement, we have a systematic way to validate our hypothesis, that is, to measure and

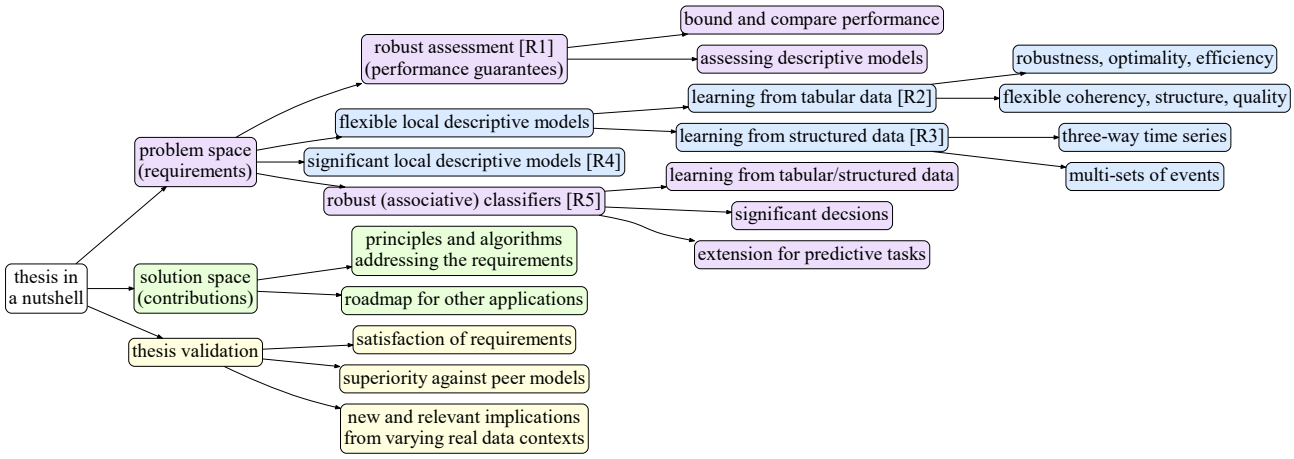

Figure 4: Structured view of the thesis scope: requirements and premises to validate the underlying hypothesis.

compare the impact of modeling regions with varying properties of interest on the target learning tasks.

R2 Learning of flexible and robust local descriptive models from tabular data.

The satisfaction of this requirement allows the systematic exploration of the impact that distinct biclustering models have in the ability to learn from high-dimensional data. This requires the scalable discovery of flexible structures of biclusters with parameterizable homogeneity criteria, yet offering optimality guarantees to properly assess their impact on descriptive and prescriptive tasks.

R3 Learning of flexible and robust local descriptive models from structured data.

Although the satisfaction of **R2** already covers different high-dimensional data contexts, such as matrices and network data, it excludes other data structures that are becoming increasingly relevant, such as multivariate time series, sequential databases and multi-sets of events. These learning challenges are thus specifically addressed under this requirement.

R4 Guarantee the statistical significance of local descriptive models.

To answer the introduced need to assess the impact of reducing dimensionality or selecting regions of interest (*Section*), we require the target local descriptive models to be statistically significant. Addressing this requirement implies the presence of a robust statistical assessment to guarantee that regions with varying coherence and quality (either from tabular or structured data) are not prone to occur by chance. This allows the inference of constraints based on the properties of these regions and the original data, that can be used to guide the learning.

R5 Learning effective classifiers from flexible, robust and statistically significant local descriptive models.

This requirement combines the previous focus on local descriptive models with the need to guarantee their discriminative power in labeled data contexts. Its satisfaction allows the assessment of the impact that the coherency, quality, significance and discriminative power of the selected regions have in the performance of classification models. The significance assessment of descriptive models is also extended for (associative) classification models and used to affect the learning. All the previous contributions are thus used at this point to guarantee both the accuracy and statistical significance of classification decisions. As a result, an integrative view of the pros and cons of learning local descriptive models to perform classification from distinct high-dimensional data domains is required.

Finally, this requirement is further extended in this work to guarantee the ability to learn from structured codomains given by sequences of classes for the adequate answering of predictive tasks.

Table 2 provides a non-exhaustive decomposition of these five structural requirements.

Given the formulation of these requirements, our work becomes a matter of testing whether they can be si-

Table 2: Decomposition of the five requirements: list of the tackled requirements.

Requirement
R1: Robust assessment of models learned from high-dimensional data; **R1.1**: Performance guarantees of classification models; **R1.2**: Performance guarantees of local descriptive models; **R1.3**: Adequate generation of synthetic data for non-biased and complete assessments;
R2: Learning biclustering models from tabular data; **R2.1**: Flexible structures of biclusters with optimality guarantees; **R2.2**: Biclustering models with varying coherency: additive, multiplicative, plaid and order-preserving models; **R2.3**: Robustness of biclustering models to: 1) different forms of noise, 2) discretization and 3) missings; **R2.4**: Scalability of biclustering searches (with optimality guarantees); **R2.5**: Extension of contributions towards network data; **R2.6**: Effective and efficient learning in the presence of background knowledge; **R2.7**: Sound integration of previous contributions;
R3: Learning local descriptive models from structured data; **R3.1**: Learning cascade models from three-way time series; **R3.2**: Learning arrangements of events from multi-sets of events; **R3.3**: Stochastic modeling of structured data;
R4: Guarantee the significance of flexible local descriptive models; **R4.1/4**: Robust assessment of the statistical significance of discrete and real-valued biclusters; **R4.2**: Robust assessment of additive, multiplicative, symmetric, order-preserving and plaid models; **R4.3/7**: Robust assessment of regions with arbitrary-high levels of noisy and missings; **R4.5**: Robust assessment of cascades and arrangements of events from structured data; **R4.6**: Learning of local descriptive models from previous statistical views;
R5: Learning accurate and significant classification models from high-dimensional data; **R5.1**: Learning effective associative classifiers from tabular data; **R5.2**: Learning effective associative classifiers from structured data contexts; **R5.3**: Learning classifiers with guarantees of statistical significance; **R5.4**: Multi-period classification: extending previous contributions for the learning of sequences of classes;

multaneously satisfied (solution space), and whether their satisfaction is associated with an improved learning in high-dimensional spaces. The thesis statement is thus asserted upon the verification of the three following sub-hypotheses: *1)* the proposed learning models satisfy the introduced requirements, *2)* these learning models offer distinctive behavior of interest against state-of-the-art learners, and *3)* their application across real-world data domains can be used to unravel new, meaningful and significant relations from data.

Solution Space

Multiple contributions resulted from addressing the introduced requirement. Contributions take two major forms: *1)* principles, and *2)* algorithms and (assessment) methodologies that rely on one or more principles. Many of these contributions are not only relevant to tackle the target problem, but can also be applied to answer other problems. In order to not compromise the line of focus of this work, the applicability of our contributions to other problems are properly identified along the text using dedicated frames (see *Notation*). As we address and answer each requirement, we expect that the proposed research produces the following scientific contributions:

C1. Methods to bound and compare the performance of local descriptors and classifiers in high-dimensional data contexts, including adequate loss functions (able to measure the impact of selecting regions with varying properties), robust error estimators and generators of data for non-biased and complete assessments;

C2. New descriptors of tabular data able to efficiently discover flexible structures of biclusters with optimality guarantees and robustness to varying forms of noise. Algorithms to retrieve non-constant coherencies, such as plaid and order-preserving models; to guarantee an adequate analysis of varying forms of tabular data (including network data); and to effectively incorporate background knowledge;

C3. Structured view on the increasingly relevant problems of learning cascades models from three-way time series and arrangements of events from multi-sets of events. Principles to handle the inherent complexity and variability of local responses in these data contexts, combining temporal and cross-attribute views with (possible) misalignments across observations. Deterministic and stochastic algorithms integrating these principles;

C4. Statistical views to robustly assess the significance of regions from tabular and structured data with regards to their coherency, quality and size (with upper limits on the risk of false discoveries). Revised algorithms to combine homogeneity (C2-C3) and significance (C4) views for guiding the learning;

C5. Principles for an adequate discovery (C2-C4), composition, scoring and testing of (informative and discriminative) regions from tabular and structured data. New associative classifiers able to incorporate previous principles. Principles to assess and promote the statistical significance of classification decisions. Systematic analysis of the performance impact of varying the properties of the underlying regions and learning functions across data domains. Extension of the proposed classifiers, preserving the accuracy and significance of the proposed learning functions (C5), to learn sequences of classes for predictive tasks.

Transversally to these set of major contributions, we additionally: 1) survey the contributions and limitations of state-of-the-art methods, and experimentally compare them against the proposed methods; and 2) show the relevance of the learned models to unravel significant and non-trivial relations across data domains, with a particular incidence on biomedical domains.

Contents

The dissertation document is organized as a set of books. Books II to VI expose the core contributions of our thesis, each book tackling one of the introduced requirements. The contents within each book are carefully discussed at their start. A book is organize in chapters. Each chapter addresses a finer requirement and delivers a compact set of contributions that become available for the following chapters and books. In this way, contents are incrementally built upon previous contents, until we are able to test the cogency of the target hypothesis.

Book II defines an assessment methodology to validate subsequent contributions. The new methods for learning flexible local descriptive models from tabular data contexts proposed in *Book III* are extended in *Book IV* towards structured data contexts, and combined in *Book V* with guarantees of statistical significance. *Book VI* proposes classifiers based on the previous models and tackles the problem of guaranteeing the statistical significance of their decisions. Finally, *Book VII* discusses the conditions on which the thesis statement is satisfied, provides an integrative view of the proposed contributions, and summarizes their major implications.

II. Performance Guarantees of Models Learned from High-Dimensional Data

Assessing the performance of descriptive and classification models is essentially a function of the selected error estimators and the properties of the input data and output model. In high-dimensional data contexts, performance is more susceptible to different sources of error. The challenges associated with learning from these data contexts were explored in previous section. As such, robustly assessing the performance guarantees of the learners is critical to validate and weight the increasingly available scientific statements derived from their application over real data. For this aim, this book proposes an assessment methodology, parameterized with robust error estimators, to validate the contributions of the following books.

Chapter 1 tackles the problem of bounding and comparing the performance of classification models learned from high-dimensional data. First, a set of prominent challenges is synthesized, including the need to adequately measure the over/underfitting propensity of the assessed models and test the statistical significance of the error estimator. Second, a set of principles is proposed to answer the identified challenges. These principles provide a roadmap of decisions to define robust statistical tests, to select adequate performance views and sampling schema, and to infer performance guarantees in the presence of multiple datasets and parameterizations of classifiers.

Chapter 2 extends these contributions towards the assessment of descriptive models learned from both tabular and structured data contexts. This an essential problem due to the lack of consensus on the applicable loss functions and the absence of principles to bound and compare the performance of descriptive models. In particular, the focus is placed on the assessment of biclustering, triclustering and cascade models due to their locality criteria, inherent flexibility and role in our thesis. As a result, this chapter surveys, compares and proposes loss functions to robustly assess these models in the presence and absence of knowledge regarding the underlying data regularities. New error estimators parameterized with these loss functions are proposed for an adequate assessment of their generalization error.

Despite the relevance of previous principles, they are insufficient to guarantee robust assessments when the performance of a learning function is evaluated over synthetic data with optimistic biases towards its behavior or from real data without a clear ground truth. In this context, *Chapter 3* proposes generators of synthetic data with parameterizable properties for an in-depth understanding of the behavior of the assessed methods. In particular, we provide generators of: matrix and network data with planted regions given by biclusters with varying structure, homogeneity and significance; multivariate time series with complex and diverse cascades; multi-sets of events with arrangements of informative events; and labeled (tabular and structured) data with global and local class-conditional regularities.

Index of Requirements and Contributions

Tables 3-5 provide an exhaustive listing of the tackled requirements and proposed contributions throughout this book. These tables aim to promote the traceability of the discussed contents per chapter and serve as an indexation and consultation tool, not dispensing the reading of their background provided by each chapter.

Table 3: Contributions to robustly assess *classification models* in high-dimensional data contexts (*Chapter* 1).

R1: Robust assessment of models learned from high-dimensional data;
R1.1: Performance guarantees of classification models;
R1.1.1: Robust statistical tests to bound and compare models;
C1.1.1a: Pairwise and multiwise comparisons with varying levels of conservatism and computational complexity;
C1.1.1b: Confidence intervals sensitive to (high) error variability, weighted by biasedness factors;
C1.1.1c: Adequate loss functions with possible smoothing factors for probabilistic models and/or outputs;
R1.1.2: Measure propensity towards under/overfitting;
C1.1.2: Decomposition of performance in bias and variance components (understand generalization);
R1.1.3: Guarantee the significance of error estimators;
C1.1.3a: Statistical tests of the feasibility of error estimates against loose settings given by null data or null models;
C1.1.3b: Adequate sampling schema (binomial tests to fix optimum training-testing split);
R1.1.4: Understand impact of data size and dimensionality;
C1.1.4a: Fitted curves for extrapolation of performance guarantees as a function of data size and dimensionality;
C1.1.4b: Parameter-dependent guarantees from the learned model (discriminant analysis and VC) and data properties;
R1.1.5: Performance guarantees from multi-data and multi-parameter assessments;
C1.1.5a: Summarization, (non-linear) regressions, visualization;
C1.1.5b: Generalized guarantees from joint analysis of estimates or learned models;
R1.1.6: Extensibility towards imbalanced data contexts;
C1.1.6a: Principles for generating data with varying properties (inc. global and local regularities and distinct sources of noise);
C1.1.6b: Adequate loss functions and estimators for imbalanced class-conditional representativity and complexity;

Table 4: Contributions to robustly assess *local descriptive models* in high-dimensional data contexts (*Chapter* 2).

R1: Robust assessment of models learned from high-dimensional data;
R1.2: Performance guarantees of descriptive models;
R1.2.1: Bounding and comparing models' performance;
R1.2.1.1: Robust estimators (collection of error estimates);
C1.2.1.1a: Sampling, randomization, peer data and multi-performance views (real data);
C1.2.1.1b: Instantiation with preserved properties (synthetic data);
R1.2.1.2: Extension of requirements R1.1-R1.7 for descriptors;
C1.2.1.2: Extensibility of C1.1 contributions (including smoothing factors for generative descriptors and feasibility tests);
R1.2.2: Adequate loss functions for (local) descriptors;
R1.2.2.1: Robust metrics for biclustering models (tabular data);
R1.2.2.1.1: Effective similarity scores (synthetic data);
C1.2.2.1.1a: Comparison of clustering views (purity, entropy, recall, precision), match scores (subspace clustering, Jaccard-based, consensus);
C1.2.2.1.1b: New integrative scores;
R1.2.2.1.2: Effective merit functions and domain-driven scores to assess relevance (real data);
C1.2.2.1.2a: Principles on how to use merit functions;
C1.2.2.1.2b: Functional enrichment tests and applicable corrections from knowledge bases, semantic sources and literature;
R1.2.2.2: Robust metrics for local descriptors learned from cube data;
R1.2.2.2.1: Effective metrics for triclustering models;
C1.2.2.2.1a: Biclustering views and extensions towards integrative cube data;
C1.2.2.2.1b: Time-sensitive similarities and merit functions;
R1.2.2.2.2: Effective metrics for cascade models (module/causality-centric and integrative);
C1.2.2.2.2a: Module-centric and causality-centric scores;
C1.2.2.2.2c: New integrative scores inspired on sequence alignments;

Table 5: Contributions on the generation of synthetic data for non-biased and complete assessments (*Chapter* 3).

R1: Robust assessment of models learned from high-dimensional data;
R1.3: Adequate synthetic data for non-biased and complete assessments;
R1.3.1: Effective generator of tabular data;
C1.3.1.1a: Structured view on the properties of biclustering models;
C1.3.1.1b: Matrix data with parameterizable size, background distributions and planted regions with varying number, shape (including extent of differences in size within a single data), positioning, coherency strength, and quality;
C1.3.1.1c: Effective generation of different forms of coherency assumption (constant, additive, multiplicative, symmetric, order-preserving) and penalization factors on their size to guarantee their statistical significance;
C1.3.1.1d: Generation of plaid structures able to model complex overlapping effects between groups of biclusters;
C1.3.1.2: Generation of network data (homogeneous or heterogeneous with weighted or labeled interactions) with parameterizable density, distributions of edges per node, distributions of interactions' score, and planted modules;
C1.3.1.3: Benchmark datasets preserving the statistical properties of experimental biological data;
R1.3.2: Effective generator of structured data;
R1.3.2.1: Generation of three-way time series;
C1.3.2.1a: Procedures to generate three-way time series with varying properties, including diverse cascades with varying distributions on the number and duration of modules, their dependencies, support and coherency;
C1.3.2.1b: Simulation of stochasticity: adequate generation of: 1) temporal and structural misalignments, and 2) bifurcations;
R1.3.2.2: Generation of multi-sets of events;
C1.3.2.2: Extension of sequential databases to plant timestamps, explore varying sparsity, and create a multiplicity of attributes;
R1.3.3: Effective generator of labeled data for classification;
C1.3.3a: Procedures to generate tabular and structured data with parameterizable number and imbalance of classes;
C1.3.3b: Combined global and local class-conditional regularities (regions effectively planted according to varying discriminative criteria);

III. Learning Local Descriptive Models from Tabular Data

The learning of flexible and robust local descriptive models, the building blocks for all subsequent contributions, is the central requirement of this work. This book tackles the task of learning local descriptive models from tabular data. These principles are then further extended to learn local descriptive models from structured data (*Book IV*) and combined with new principles to guarantee the statistical significance of these models (*Book V*). As a result, these local descriptive models provide the necessary basis to learn associative classifiers and study their behavior (*Book VI*). As such, in order to understand how the number, positioning, size, coherence and quality of the selected regions affect the performance of both descriptive and classification models: we need to guarantee the *flexibility* and *robustness* of the target learners. Furthermore, *optimality* guarantees need to be present for a reliable quantification of the impact of the selected regions in the performance of descriptive and classification models, as well as *scalability* guarantees to enable the learning from high-dimensional tabular data.

In this context, the biclustering task, aiming to discover subsets of features exhibiting a coherent pattern over a subset of observations, is by definition able to achieve these qualities. In fact, biomedical and social data are characterized by the presence of local regions, whose discovery has been largely studied and motivated in the context of biclustering [19]. However, due to the complexity of the biclustering task, most of the existing biclustering algorithms place restrictions on the coherency, quality and structure of biclusters (preventing the recovery of all relevant regions) and/or rely on greedy or stochastic approaches (associated with suboptimal solutions) [20]. To address this problem, this book proposes a new class of biclustering models. As such, to guarantee the flexibility, optimality, robustness and scalability of this learning task, the following four major requirements need to be satisfied:

- flexible structures of biclustering with guarantees of optimality **[R2.1]**;
- biclusters with flexible (yet meaningful) coherency **[R2.2]**;
- biclusters robust to different types and amount of noise **[R2.3]**;
- scalable searches in flexible settings (loose constraints) **[R2.4]**.

Furthermore, additional requirements are placed to learn local models from network and tabular data [**R2.5**], and to effectively guide the learning in the presence of background knowledge and user expectations [**R2.6**]. These six requirements are depicted in Figure 5 and further decomposed in the provided tables throughout this section.

A new class of biclustering approaches based on principles from pattern mining [19, 28, 27], referred in this work as pattern-based biclustering, is well-positioned to answer the introduced requirements as it is able to exhaustively (yet efficiently) discover flexible structures of biclusters. However, state-of-the-art contributions are dispersed and the potential of their integration remains unclear. *Chapter 1* provides a structured view on pattern-based biclustering. For this purpose, we survey its potentialities and limitations, and make available a set of principles for a guided definition of new pattern-based biclustering approaches that are able to combine existing contributions as well as accommodate new contributions.

Chapter 2 proposes new principles to guarantee the robustness of biclustering solutions to discretization procedures, varying forms and amount of noise, and arbitrary high-levels of missing values. These principles are integrated within a new pattern-based biclustering approach, referred as BicPAM (Biclustering based on PAttern Mining), to allow the discovery of biclusters with robust yet (possibly) parameterizable quality.

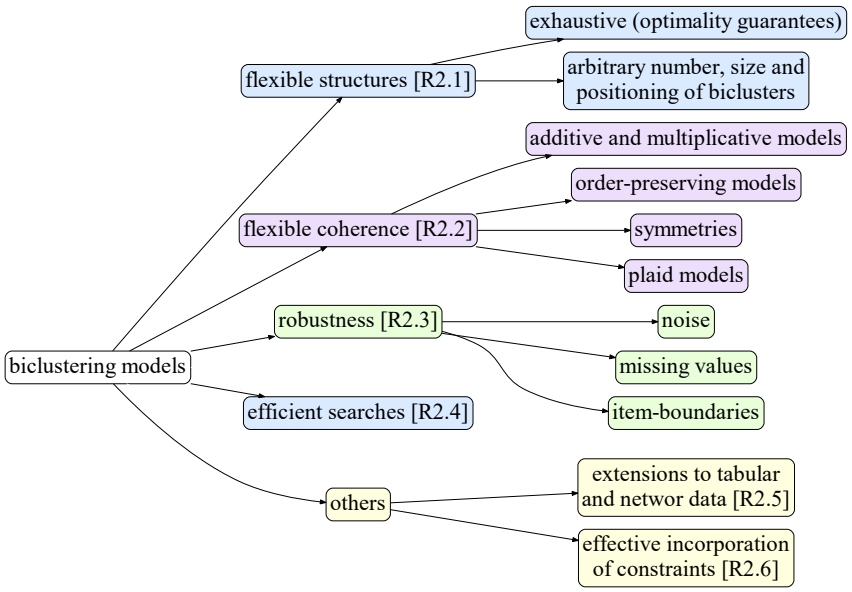

Figure 5: Requirements associated with the target task of learning flexible biclustering models.

Chapter 3 extends the constant assumption of pattern-based biclustering towards more flexible coherencies, including additive and multiplicative coherencies in the presence or absence of symmetries. In this context, we first motivate their need across biomedical and social data and provide principles for their exhaustive discovery that surpasses the limitations of existing searches. Second, we extend the biclustering task towards tabular data with non-identically distributed features.

Chapter 4 enhances the previous searches with principles to model biclustering solutions with a plaid model. The plaid model considers a cumulative composition of contributions in the areas where the biclusters overlap, thus being required to accommodate meaningful interactions between biclusters in biological and social domains. We propose BiP (Biclustering using Plaid models), an extension of BicPAM for recovering excluded regions due to unaccounted plaid effects. To address prominent restrictions associated with the plaid model, we further propose meaningful relaxations to allow a noise-tolerant composition of contributions and integrate them in BiP.

Chapter 5 tackles the efficiency bottlenecks of pattern-based biclustering searches. First, a new search based on annotated pattern-growth trees is proposed to avoid the use of expensive data structures and procedures required to maintain the sets of observations and features per region of interest. Second, we introduce principles to foster the scalability of the learning methods for very large data while still providing optimality guarantees, including searches in distributed and partitioned data settings or targeting biclusters derived from approximative patterns.

Chapter 6 explores a last form of coherency commonly observed in biomedical and social data domains: the order-preserving coherency. The values of the observations of an order-preserving bicluster induce the same linear ordering across features. This chapter tackles the limitations of existing order-preserving approaches, which either suffer from scalability or flexibility issues and are not able to discover biclusters with symmetries and parameterizable noise-tolerance. For this purpose, we first propose new searches able to seize efficiency gains from the item-indexable properties associated with the order-preserving biclustering task. Second, these searches are enhanced with new principles to guarantee robustness to noise and to accommodate symmetries.

The previous chapters introduce scalable searches for pattern-based biclustering under the assumption that the postprocessing step is accomplished efficiently. However, in the presence of voluminous biclustering solutions, the cost of computing similarities between all pairs of biclusters or testing the inclusion/removal of observations and features per bicluster can be highly expensive. *Chapter 7* provides principles to guarantee the efficiency of postprocessing procedures.

Chapter 8 extends the previous contributions towards network data. We propose BicNET (Biclustering NET-

works), an algorithm to discover non-trivial yet coherent modules in weighted graphs with heightened efficiency. First, we motivate the relevance of discovering network modules given by biclusters with flexible coherencies and tolerance to missing and noisy interactions in order to guarantee a focus on meaningful yet non-necessarily dense modules. Second, we adapt the underlying data structures and searches to seize high efficiency gains from the inherent structural sparsity of network data.

Chapter 9 combines all previous contributions. In this context, we provide the integrative algorithmic solution, analyze its computational complexity, and describe how its behavior can be either dynamically or parametrically customized based on the properties of the input data.

Chapter 10 extends pattern-based biclustering to effectively seize efficiency gains in the presence of constraints. In particular, we illustrate how constraints with succinct, (anti-)monotone and convertible properties can be derived from knowledge repositories and user expectations, and discuss and how they can be effectively used to prune the search space and guarantee a focus on regions of interest.

Index of Requirements and Contributions

Tables 6-15 provide an enumeration of the proposed contributions throughout the chapters of this book.

Table 6: Major contributions for the exhaustive discovery of flexible structures of biclusters (*Chapter* 1).

R2: Efficient learning of biclustering models with flexible structures, coherency and quality;
R2.1: Discovery of flexible structures of biclusters with optimality guarantees;
C2.1.1a: Formal mapping of biclustering as frequent itemset mining and association rule mining;
C2.1.1b: Principles for declarative biclustering based on constraint-based mining and formal concepts;
C2.1.1c: Formal mapping of biclusterings as structured pattern mining, including sequential pattern mining and graph mining;
C2.1.2: Comparison of pattern representations (with impact on structures), including simple, maximal and closed;
C2.1.3: Preprocessing: principles to dynamically select adequate normalization and discretization procedures, and to handle outliers;
C2.1.4: Survey of alternative (anti-)monotonic for learning directly from real-valued data (with impact on structures);
C2.1.5: Adequate postprocessing options to affect the properties of the target structures;
C2.1.6: Systemic (qualitative and quantitative) comparison of existing pattern-based biclustering algorithms;
C2.1.7: Structured view of pattern-based biclustering: guidelines for the development of new algorithms;
R2.2: Ability to learn biclustering models with varying coherency assumptions;
R2.3: Biclustering solutions robust to noise and parameterizable quality;
R2.4: Efficiency of biclustering searches (optimality guarantees);
R2.5: Extension of original contributions towards tabular and network data;
R2.6: Effective learning in the presence of background knowledge;
R2.7: Sound integration of contributions;

Table 7: Major contributions to guarantee the robustness of biclustering models (*Chapter* 2).

R2.3: Biclustering solutions robust to noise and parameterizable quality;
R2.3.1: No exposure to the items-boundary problem (discretization drawback);
C2.3.1: Effective principles to handle item-boundaries based on multi-item assignments;
R2.3.2: Efficient discovery of robust biclusters with calibrated type and amount of noise;
C2.3.2.1: Noise-tolerance through the use of association rules with parameterizable confidence and interestingness;
C2.3.2.2: Calibrating coherency strength through mapping options (alphabet, normalization and discretization);
C2.3.2.3: Guaranteeing robustness through adequate closing options;
C2.3.2.3a: Principles for effective merging and filtering of biclusters;
C2.3.2.3b: Principles for effective extension and reduction of biclusters;
R2.3.3: Adequate handling of missing values;
C2.3.3a: Strict handling of missings: removal, replacement as dedicated symbol, and imputation;
C2.3.3b: Relaxed handling of missings: multi-item assignments based on imputed value;

Table 8: Contributions for learning additive, multiplicative and symmetric models (*Chapter* 3).

R2.2: Ability to learn biclustering models with varying coherency assumptions;
R2.2.1: Learning additive and multiplicative models;
C2.2.1.1: Exhaustive discovery of additive models based on shifting factors per observation or feature;
C2.2.1.2: Exhaustive discovery of multiplicative models based on common multiples;
C2.2.1.3: Principles for pruning the search space and to accommodate noise relaxations;
R2.2.2: Learning plaid models;
R2.2.3: Learning order-preserving models;
R2.2.4: Learning coherent models in the presence symmetries;
C2.2.4a: Exhaustive accommodation symmetries over constant models and adequate pruning of the search space;
C2.2.4b: Extensions for testing symmetries within additive models;
R2.5: Extension of original contributions towards new data contexts;
R2.5.1: Learning from network data;
R2.5.2: Learning from tabular data;
C2.5.2.1: Adequate procedures to discretize and balance feature's domains for the application of pattern-based biclustering;
C2.5.2.2: Principles to address limitations associated imbalanced feature's domains;
C2.5.2.2: Method to find biclusters with mixtures of coherency assumptions;

Table 9: Major contributions for the effective learning of plaid models (*Chapter* 4).

R2.2.2: Learning plaid models;
R2.2.2.1: Effective learning of complex plaid models;
C2.2.2.1.1: Principles to recover excluded areas due to plaid effects and to exclude noisy areas non-explained by plaid effects;
C2.2.2.1.2: Residue-based heuristic for learning complex plaid models (large sets of interacting biclusters with non-trivial overlaps);
C2.2.2.2.3: Incorporation of previous contributions to allow the discovery of plaid models with non-constant coherencies and not requiring an exhaustive coverage of all data elements;
C2.2.2.2.4: Systematization of possible interactions between biclusters across domains (e.g. is-part-of, depends-on);
R2.2.2.2: Addressing the restrictive exact additive match;
C2.2.4.2.1: Definition and efficient inclusion of meaningful relaxations: a) in-between and b) noise-tolerant matches;
C2.2.4.2.2: Efficient inclusion of advanced composition functions: a) weighted and b) scaling composition functions;

Table 10: Proposed contributions to guarantee the efficiency of pattern-based biclustering (*Chapter* 5).

R2.4: Efficiency of biclustering searches (optimality guarantees);
R2.4.1: Searches less susceptible to current time-memory bottlenecks;
C2.4.1.1: Comparison of full-pattern mining searches for biclustering;
C2.4.1.2: Extended pattern tree (FP-tree) with transactions annotations (minimum overhead) to tackle costs associated with bitsets;
C2.4.1.3: New pattern-growth algorithm with efficient build and traversal of annotated FP-trees for efficient pattern-based biclustering on dense and high-dimensional data;
C2.4.1.4: Principles to customize search options from data size, dimensionality and regularities;
R2.4.2: Efficient sequential pattern mining in item-indexable data;
R2.4.3: Efficient postprocessing (merging, extension and reduction procedures);
R2.4.4: Scalability of pattern mining searches;
C2.4.4.1: Data partitioning principles with constraint-guided searches to preserve optimality;
C2.4.4.2: Parallelization principles and mining of approximative patterns (with optimality guarantees);

Table 11: Contributions for the robust learning of order-preserving models in the absence and presence of symmetries (*Chapter* 6).

R2.2.3: Learning order-preserving models;
R2.2.3.1: Efficient and exhaustive discovery of order-preserving models;
C2.2.3.1.1: Pattern-based biclustering with sequential pattern mining over mapped sequential databases from tabular data;
C2.2.3.1.2: Formulation of the new task of SPM over item-indexable databases (sequential database from ordered tabular data);
C2.2.3.1.3: New SPM pattern-growth search based on efficient data-projections and compact data structures;
C2.2.3.1.4: Principles for pruning search space in the presence of user expectations;
R2.2.3.2: Flexible order-preserving models;
C2.2.3.2.1: Methods to mine both monotonically and strictly increasing orders;
C2.2.3.2.2: Methods for the exhaustive discovery of order-preserving models with symmetries;
C2.2.3.2.3: Parameterizable degree of co-occurrences versus precedences according to the target task and data properties;
R2.2.3.3: Robustness of order-preserving models;
C2.2.3.3.1: Inclusion of previous principles to handle varying levels of noise and missings;
C2.2.3.3.2: Compliance of order-preserving searches with preprocessing and postprocessing options;

Table 12: Proposed methods to guarantee the efficiency of merging, extension and reduction procedures (*Chapter* 7).

R2.4.3: Efficient postprocessing options;
R2.4.3.1: Efficient merging procedures;
C2.4.3.1.1: Mapping merging as the task of discovering maximal circuits in unweighted undirected graphs;
C2.4.3.1.2a: Formulation of frequent itemset mining task with variable support (multi-support FIM);
C2.4.3.1.2b: Mapping merging into multi-support FIM;
C2.4.3.1.2c: Efficient methods for multi-support FIM and implications;
C2.4.3.1.3: Anti-monotonic heuristics for efficient calculus of similarities (with short-circuiting verifications);
C2.4.3.1.4: Pushing merging procedures to the mining step (similarities using operations over tree structures);
R2.4.3.1: Efficient extension and reduction procedures;
C2.4.3.1a: Pattern-based biclustering with varying support and pattern length thresholds to guide extensions and reductions;
C2.4.3.1b: Constraint-based searches over specific regions of interest to guarantee heightened postprocessing efficiency;

Table 13: Contributions for the efficient and robust biclustering of network data to find non-trivial (yet coherent) modules (*Chapter* 8).

R2.5.1: Learning from network data;
R2.5.1.1: Efficiency of biclustering methods for large-scale networks;
C2.5.1.1.1: Extension of pattern-based biclustering to learn from sparse data;
C2.5.1.1.2: Efficient data structures and principles for adequate space exploration;
R2.5.1.2: Discovery of network modules with non-dense yet meaningful coherency;
C2.5.1.2.1: Incorporation of previous principles to discover modules/biclusters with flexible coherency;
C2.5.1.2.1a: Constant and symmetric models for finding modules with non-necessarily high (yet coherent) interactions;
C2.5.1.2.1b: Plaid models to accommodate weight variations associated with network topology (hubs, and between- and within-pathway interactions;
C2.5.1.2.1c: Order-preserving models for discovering modules with coherent degree of influence between sets of nodes;
C2.5.1.2.2: Extension of principles for different types of networks;
C2.5.1.2.2a: Principles for biclustering homogeneous and heterogeneous networks;
C2.5.1.2.2b: Principles for biclustering networks with quantitative (weights) and qualitative (labels) interactions;
R2.5.1.3: Discovery of robust network modules;
C2.5.1.3a: Principles for biclustering modules robust to missing interactions (tackling fully-interconnectedness restriction);
C2.5.1.3b: Principles for biclustering modules robust to noisy interactions;

Table 14: Consistent integration of previous contributions and exploration of their synergies (*Chapter* 9).

R2.7: Sound integration of previous contributions;
C2.7.1a: Exploration of efficiency gains from synergies associated with the integrative search for different coherency assumption;
C2.7.1b: Efficiency gains from applying biclustering with varying levels of tolerance to noise, levels of coherency strength and from the combined application of preprocessing and postprocessing steps;
C2.7.2: Consistent algorithmic basis and the analysis of its computational complexity;
C2.7.3: Robust default and dynamically parameterized behavior based on the input data properties;
C2.7.4a: Structured view of the parameters of pattern-based biclustering to customize the coherency, quality and structure of solutions;
R2.7.4b: Declarative interface for the constraint-based definition of the desirable properties of biclustering solutions;
C2.7.4c: Graphical interface for usable parameterization, display of results and soundness checks;
C2.7.4d: Programmatic interface to flexibly control, extend and adapt the biclustering behavior;
R2.5.3: Effective learning from sparse data;
C2.5.3: Principles for biclustering data with an arbitrary-high number of uninformative elements and/or missings;

Table 15: Contributions for biclustering in the presence of background knowledge (*Chapter* 10).

R2.6: Effective learning in the presence of background knowledge;
R2.6.1: Effective incorporation of constraints with nice properties;
C2.6.1.1a: Structured view on the relevance and properties of constraints with succinct, (anti-)monotonic and convertible properties for biclusters from frequent itemsets for matrix and network data;
C2.6.1.1b: Extensibility of constraints for plaid models from association rules;
C2.6.1.1c: Prefix-monotone constraints and regular expressions for order-preserving models from sequential patterns;
C2.6.1.2: Comparison of efficiency gains from available principles for domain-driven pattern mining;
C2.6.1.3a: Compliance of F2G searches for pattern-based biclustering with CFG principles;
C2.6.1.3b: Compliance of F2G searches for with Bonsai and data-reduction principles;
C2.6.1.4: Extension of IndexSpan searches for with prefix-monotone verifications during database projections;
R2.6.2: Guided instantiation of nice constraints;
C2.6.2.1: Roadmap with illustrative sets of succinct, (anti-)monotonic and convertible across biomedical domains (expression data and biological networks) and social domains;
C2.6.2.2: Formalism and procedures to interpret (parametric) constraints to customize the biclustering behavior with sharp usability;
R2.6.3: Effective learning in the presence of background knowledge;
R2.6.3: Effective learning in the presence of knowledge-driven annotations;
C2.6.3: Guided searches when rows and columns are annotated with labels extracted from knowledge bases and literature;

IV. Learning Local Descriptive Models from Structured Data

The previous book, *Book III*, tackled the task of learning flexible local descriptive models from tabular data with guarantees of flexibility, robustness and efficiency. However, as we move from tabular to structured data, new learning functions need to be defined. Multivariate time series are increasingly collected in biological and social domains to study regulatory and behavioral responses. Data from public and private sectors typically follow multi-dimensional or relational schema. In this context, learning from multi-sets of events is gaining momentum for the analysis of repositories of health-records, user actions or financial transactions. These observations stress the importance of adequately learning from these data contexts. However, the state-of-the-art contributions are challenged by their high-dimensionality and the need to model both structural and temporal relations. This book extends our research scope with principles to effectively and efficiently learn local descriptive models from structured data, including multivariate time series and multi-sets of events.

Chapter 1 tackles the task of learning cascade models from three-way time series. Learning from these data contexts is challenged by their stochasticity and complexity associated with the variability of temporal dynamics across observations and with the probabilistic elicitation of distinct paths within a cascade. Addressing these challenges is critical to model regulatory responses to drugs, disease progression, growth and development, as well as behavioral cascades (associated with social interaction, web navigation, commercial activity, financial decisions) to specific events of interest. In this context, this chapter provides structured view of the problem and new algorithms to model the structural and temporal kinetics of responses from three-way time series.

Chapter 2 addresses the task of learning arrangements of informative events from multi-sets of events. Learning from multi-sets of events is challenged by the structural sparsity of event occurrences (arbitrary temporal distances between events) and by the presence of distinct types of events (multiplicity of attributes capturing different views of interest). Despite the criticality of learning integrative models able to simultaneously capture temporal dynamics across distinct types of events, state-of-the-art contributions fail to satisfy this essential requirement. As such, this chapter explores new data structures and learning functions to robustly and efficiently model informative groups of events related through temporal and cross-attribute dependencies. Furthermore, it guarantees its extensibility towards highly structured data given by relational and multi-dimensional schema, as well as by less structured data associated with repositories of (health/action/decision) records.

In *Chapters 3* and *4*, the problem of modeling local regions of interest (characterized by inherent temporal and structural relations) from three-way time series and multi-sets of events is tackled by mapping data into (time-enriched) itemset sequences and deterministically exploring them. Despite the benefits of the proposed deterministic searches, they are associated with large outputs (computationally expensive postprocessing procedures), uneven space exploration in the presence of a few larger regions, and efficiency bottlenecks for highly dense data. Despite the relevance of stochastic searches to address these problems, existing contributions in the context of dynamic Bayesian models, neural networks and Markov-based models are only prepared to deal with sequences with a fixed multivariate order and to model patterns with well-known shape. As such, these solutions can be seen as prototypical and clearly insufficient to tackle the target task. In this context, *Chapter 3* proposes contributions to stochastically model itemset sequences under a probabilistic and noise-tolerant view of frequent orderings, offering the possibility to focus the search on dissimilar and arbitrary-large regions. For this aim, we propose new classes of hidden Markov models to learn a compact and generative representation of patterns combining both co-

occurrences and precedences, exploring new architectures, convergence schema and graph propagation procedures to guarantee a robust learning. Finally, a comprehensive comparison of probabilistic and deterministic functions is provided.

Finally, *Chapter 4* extends the contributions proposed in the context of the previous chapter to address specific challenges associated with the (stochastic) modeling of three-way time series and multi-sets of events. The learning of cascades from three-way time series is challenged by the complexity and diversity of cascades, misalignments across observations, and high-dimensionality. The learning of arrangements from multi-sets of events is challenged by the arbitrary sparsity of events and the relevance of framing their occurring time. In this context, the proposed hidden Markov models are extended to deal with the inherent complexity of the underlying responses, handle numeric inputs and enriched to decode temporal distances between modules/events.

The principles proposed throughout this book are combined in *Book V* (*Chapter 4*) with additional principles to guarantee the statistical significance of cascades and arrangements of events. As a result, these regions provide the necessary guarantees of homogeneity and statistical significance that enable an effective learning of associative classifiers from high-dimensional structured data (*Book VI*).

Index of Requirements and Contributions

Tables 16-19 exhaustively list the proposed contributions throughout this book.

Table 16: Major contributions associated with the learning of cascade models from three-way time series (*Chapter* 1).

R3: Learning local descriptive models from structured data;
R3.1: Learning cascade models from three-way time series;
C3.1a: Formal view of the learning task and specification of its requirements;
C3.1b: List of applications and structured view on the contributions and shortcomings of related research streams;
R3.1.1: Dealing multiple observations;
R3.1.1.1/2: Modeling temporal misalignments (starting time, duration and distance between modules) across observations;
C3.1.1a: Mapping the learning of cascades as the task of modeling of frequent orderings from itemset sequences (able to accommodate arbitrary-high temporal misalignments from stochastic uncertainties);
C3.1.1b: Robust preprocessing and data mapping procedures (including multi-item assignments to surpass discretization problems);
C3.1.1c: Deterministic algorithm for the discovery of frequent co-occurrences (modules) and precedences (causal relations);
R3.1.2: Handling the diversity and complexity of responses;
R3.1.2.1: Modeling multiple responses with (possibly) multiple divergent paths;
C3.1.2a: Constraint-guided discovery of bifurcated paths with lower support from (possibly) incomplete cascades;
C3.1.2b: Postprocessing procedures based on the similarity between cascades (using alignment scores for merging and filtering) to guarantee the modeling of multiple (yet dissimilar) responses;
R3.1.2.2: Guaranteeing a balance between representativity, size, similarity and non-triviality;
C3.1.2c: Parameterization of the mapped task with dedicated algorithms to mine k-dissimilar cascades and compressed representations;
C3.1.2d: Extension and reduction of cascades based on the observed homogeneity;
C3.1.2e: Effective incorporation of constraints to focus the search on non-trivial cascades;
R3.1.3: Flexible modeling of modules;
R3.1.3.1/3: Modeling flexible structures of modules with flexible coherence;
C3.1.3a: The learning of cascades as a SPM task allows for a parameterizable coherency strength, flexible structures (possibly manipulated during the mining and postprocessing stages) and modules with varying values across time;
R3.1.3.2: Allowance for temporal shifts and scales;
C3.1.3b: Learning from multiple temporal granularities and presence of other strategies (including extension and merging procedures, and multi-item assignments) to handle structural misalignments (time shifts and scales);
R3.1.4: Discovering causality: validating dependencies between modules and investigating multiple forms of dependency;
C3.1.4a: Post-analysis procedure to differentiate causal from parallel relations between modules;
C3.1.4b: Principles to test the causality of the inferred dependencies in biological settings from transcriptional analysis;
R3.1.5: Guarantee the efficiency of the learning;
C3.1.5a: Principles to enhance scalability: 1) vertical data formats, 2) space exploration tuned to handle lengthy sets of co-occurrences, 3) bounded space searches using approximative and top-K dissimilar patterns, 4) data partitioning and parallelization principles;
C3.1.5b: Removal of non-interesting elements from the input data (increased data sparsity);
C3.1.5c: Guided discovery from user expectations: minimum number of items per module, modules per cascade or items per cascade;

Table 17: Proposed contributions for learning arrangements of events from multi-sets of events (*Chapter* 2).

R3.2: Learning arrangements of events from multi-sets of events;
C3.2a Structured view on the relevance, requirements and applicability of this task;
C3.2b Survey of contributions and limitations of integrative learning (at the input, model and output level) to model multi-sets;
R3.2.1: Modeling arrangements of events with cross-attribute dependencies (attribute multiplicity);
C3.2.1a: Mapping of multi-set of events into integrative (and time-enriched) itemset sequence conducive to the learning of arrangements of events from distinct attributes;
C3.2.1b: Principles to combine numeric/ordinal/categoric types of events with imbalanced domain cardinality;
R3.2.2: Modeling temporal dependencies among informative events with arbitrary levels of sparsity;
C3.2.2.1a: Time-enriched itemset sequences able to preserve time distances between events from different attributes;
C3.2.2.1b: Retrieval of arrangements of events through a new class temporal patterns based on the deterministic discovery of frequent orderings with time guarantees (time-sensitive sequential pattern mining);
C3.2.2.2a: Principles to effectively and efficiently learn from multiple temporal granularities to handle: 1) the structural sparsity of events, and 2) mismatched orderings associated with noise and deviations on the time occurrence of events;
C3.2.2.2b: Penalization schema so orderings from coarser-grained partitions do not jeopardize the learning;
C3.2.2.3: Adaptation of SPM methods to accommodate efficient data structures (hash structures and pointers) and searches able to preserve temporal information to frame the expected time of occurrence of the events within an arrangement;
C3.2.2.4: Adequate postprocessing to guarantee the discovery of a compact number of arrangements with dissimilarity guarantees;
R3.2.3: Learning from varying data structures;
R3.2.3.1: Handling multi-dimensional and relational databases;
C3.2.3.1a: Principles to handle: 1) memory bottlenecks associated from fact tables with high number of measures/high-dimensionality (using sparse representations), and 2) attributes with structured domains (using integrative feature extraction);
C3.2.3.1b: Mapping of relational into multi-dimensional databases;
R3.2.3.2: Handling collections of events and repositories of records;
C3.2.3.2a: Mapping of collections into multi-sets of events, and of multi-sets of events into temporal structures conducive to learning;
C3.2.3.2b: Structured view on specific contributions for learning patterns from large-scale repositories of (health) records;

Table 18: Partial list of contributions associated with the stochastic modeling of sequential patterns (*Chapter* 3).

R3.3: Stochastic modeling of structured data;
C3.3a: Structured view on the limitations of deterministic learning functions and on the opportunities of probabilistic alternatives;
R3.3.1: Stochastic modeling of itemset sequences (precondition for the analysis of other structures);
C3.3.1a: Limitations of existing stochastic methods with regards to the allowed data structures, pattern types and convergence schema;
C3.3.1b: Comparison of deterministic and probabilistic outputs w.r.t. their completeness, correctness, efficiency and intrinsic benefits;
R3.3.1.1: Modeling of frequent precedences and co-occurrences sensitive to noise;
C3.3.1.1.1a: Extension of different classes of hidden Markov models (HMMs) to learn sequential patterns by: mapping itemset sequences into univariate sequences and revising the learning to adequately distinct co-occurrences from precedences;
C3.3.1.1.1b: New HMM architectural components more prepared to model sequential patterns with varying number of precedences (by introducing deletion and skip states between itemset-aligning states) and co-occurrences;
C3.3.1.1.1c: Principles for the effective initialization of transitions and emissions of the proposed architectures from data expectations;
C3.3.1.1.2: Graph propagation procedures and cut-circuit heuristics for the effective and efficient decoding of sequential patterns from the most probable paths of the learned lattices;
R3.3.1.2: Surpassing convergence problems and robustness to spurious background matches;
C3.3.1.2a: Multi-path architectures and/or iterative learning-decoding-masking;
C3.3.1.2b: Learning settings that guarantee an adequate convergence of emissions (neither too loose such as given by different forms of expectation-maximization nor too strict such as given by entropy-based priors);
R3.3.1.3: Modeling of dissimilar, lengthy and non-trivial patterns;
C3.3.1.3a: Integrative architectures to model (a possibly high number of) dissimilar sequential patterns;
C3.3.1.3b: Effective incorporation of user expectations and background knowledge (through parameterizable architectures);
R3.3.2: Stochastic modeling of three-way time series;
R3.3.3: Stochastic modeling of multi-sets of events;

Table 19: Major contributions for the stochastic learning of local descriptive models from structured data (*Chapter* 4).

R3.3.2: Stochastic modeling of three-way time series;
C3.3.2.1: Evidence that the proposed HMMs can handle temporal misalignments and model dissimilar (possibly bifurcated) cascades;
C3.3.2.2: New mapping and extension of the proposed architectures to correctly interpret elements with multiple items assigned;
C3.3.2.3: Principles to learn from time series with fixed order (in the absence of multi-item assignments and removal of elements);
C3.3.2.4a: Adaptation of HMMs to learn directly from numeric inputs (continuous HMMs), including principles to guarantee an adequate coverage of patterns and surpass spurious background matches;
C3.3.2.4b: Structured view on the benefits and limitations, and on when and how to apply continuous HMMs;
C3.3.2.5a: Postprocessing to surpass rigidity associated with the parameterization of architectures parameterizations;
C3.3.2.5b: Advanced aspects of stochastic learning of cascades: 1) incremental learning (iterative adjustments), 2) separation of parallel from causal relations, and 3) modeling of plaid effects;
R3.3.3: Stochastic modeling of multi-sets of events;
C3.3.3.1a: Evidence of the adequacy of the proposed HMMs to model arrangements of events from time-enriched itemset sequences;
C3.3.3.1b: Discard of uninformative events from large-scale data through skip paths and calibrated removal of less probable emissions;
C3.3.3.1c: Postprocessing procedures to guarantee the adequacy of outputs when learning from multiple time partitions;
C3.3.3.2: Revised initialization of transition-emission probabilities to deal with the structural sparsity of multi-sets of events;
C3.3.3.3: Proposal of the first class of temporally-enriched HMMs: extended architectures and learning-and-decoding principles to provide temporal guarantees associated with the occurrence of events within an arrangement;

V. Significance Guarantees of Local Descriptive Models

Assessing the statistical significance of the local (descriptive) models data is required to guarantee the discovery of relevant regions from high-dimensional data, as well as to filter, validate or weight the increasing number of implications in literature derived from the analysis of biomedical and social data. As largely motivated, guaranteeing the statistical significance of the modeled regions of interest is required to minimize the propensity of a given learning function to over/underfit the observed data. This ensures the adequacy of the capacity term of a learning function, and thus its ability to generalize.

Despite the relevance of guaranteeing the statistical significance of local descriptive models, there is not yet a ground truth on how to assess the significance of flexible biclustering models from tabular data neither of cascade and event-set models from structured data. As such, this book proposes principles for the robust assessment of regions with varying criteria of homogeneity (including varying coherency and tolerance to noise) learned from tabular and structured data contexts.

Furthermore, this book integrates the proposed significance views with homogeneity views to affect the learning of local models. To illustrate this requirement, consider the problems associated with two major learning options. A first option (with propensity to overfit data) is to model regions with high homogeneity only. However, optimizing homogeneity levels is of limited use since good levels can appear by chance for small regions. A second option (with propensity to underfit data) is to model large regions (satisfying some homegeneity criteria) to minimize the chance of delivering non-significant biclusters. For this aim, some loose form of coherency is assumed and high levels of noise are tolerated. Understandably, these classes can be seen as distinct poles of the performance axis. The first option neglects the significance component of performance, while the second option prioritizes significance in detriment of homogeneity (coherency and quality).

The need to guarantee the significance of local descriptive models can be decomposed according to seven major requirements:

- robust statistical tests for biclustering, with an efficient assessment of deviation from expectations **[R4.1]**;
- significance assessment of additive, multiplicative, symmetric, order-preserving and plaid models **[R4.2]**;
- significance assessment of biclusters with arbitrary-high levels of noise **[R4.3]**;
- significance assessment of biclusters discovered in real-valued data contexts and of biclusters with continuous ranges of shifting and scaling factors **[R4.4]**;
- significance assessment of cascade models from three-way time series and of arrangements of events from multi-sets of events **[R4.5]**;
- inference of global constraints to enforce bounded guarantees on the significance of regions, and their effective incorporation within the learning process **[R4.6]**;
- combined homogeneity-significance views for adequately assessing the relevance of regions **[R4.7]**.

Chapters 1-3 propose methods to test and ensure the statistical significance of regions in tabular data contexts.

Chapter 1 motivates this need and surveys the current limitations that prevent the assessment of flexible biclustering models. To address these limitations, we propose a robust statistical framework to: 1) assess biclusters using stochastic and frequentist views able to adequately model the impact of data dimensionality; 2) minimize the number of false positives (outputted non-significant biclusters) and false negatives (non-retrieved significant

biclusters) by effectively and efficiently testing deviations from expectations; and 3) infer global constraints that guarantee the significance of a given bicluster when they are satisfied.

Chapter 2 provides the first attempt for the statistical assessment of biclusters with flexible coherency and quality by extending the previous contributions towards biclustering models with non-constant coherency assumptions and arbitrary-high levels of noise, and consistently integrating quality and significance views.

Chapter 3 extends the previous contributions towards real-valued data contexts where the coherency of a given bicluster may not be known apriori. Principles from integral calculus are used to enlarge the proposed statistical assessments to biclusters characterized by continuous ajustement factors. Finally, we guarantee the applicability of the proposed contributions for data domains with non-identically distributed features.

To guarantee the consistency of the contents in this book with the standard notation in statistics (instead of machine learning), we revised the notation of two concepts. Given a tabular dataset $\mathbf{A}$ with $\mathbf{X}$ observations and $\mathbf{Y}$ features, and a bicluster $\mathbf{B}=(\mathbf{I}\subseteq\mathbf{X},\mathbf{J}\subseteq\mathbf{Y})$, then $N=|\mathbf{X}|$ is the data size, $M=|\mathbf{Y}|$ is the data dimensionality, $n=|\mathbf{I}|$ is the number of observations in $\mathbf{B}$ and $m=|\mathbf{J}|$ is the number of features in $\mathbf{B}$.

Finally, *Chapter 4* provides principles to test and ensure the statistical significance of regions in structured data contexts. For this aim, we enrich existing statistical views of sequential data to statistically assess cascade models learned from three-way time series and arrangements of events from multi-sets of events.

Index of Requirements and Contributions

Tables 20-23 exhaustively list the proposed contributions throughout this book.

Table 20: Major contributions to assess the statistical significance of biclustering models (*Chapter* 1).

R4: Guarantee the significance of flexible local models;
R4.1: Robust assessment of the statistical significance of (discrete) biclusters;
- **C4.1a**: Structured view on the contributions/limitations to statistically model biclusters from PM, biclustering and inferential statistics;
- **C4.1b**: Systemic empirical analysis on how bicluster's properties (support, length, pattern, coherency, quality) and data properties (size, dimensionality, regularities) affect statistical significance;

R4.1.1: Robust statistical tests;
- **C4.1.1.1**: Statistical tests to assess significance of biclusters based on bionomial tails estimated from: 1) the observed regularities (using either stochastic or frequentist views), 2) generated data, and 3) hold-out partitions;
- **C4.1.1.2**: Extension of the proposed tests to guarantee their applicability to biclusters with different forms of constant coherency;
- **C4.1.1.3a**: Extension of bionomial calculus to adequately measure the impact of data dimensionality on the significance;
- **C4.1.1.3b**: Removal of efficiency bottlenecks when assessing bicluster's patterns with non-indexed columns;
- **C4.1.1.4a**: Principles to compute sound statistical views when assuming varying forms of dependency (inc. pairwise and overall) based on: 1) joint probability calculus, and 2) generated data based on n-wised dependencies;
- **C4.1.1.4b**: Dynamic programming principles to tackle efficiency bottlenecks when testing biclusters with n-wised dependencies;
- **C4.1.1.5**: Integrative assessment method where statistical decisions (tests, stochastic/frequentist views, dependency form) are dynamically selected from the input data properties or user expectations;

R4.1.2: Non-conservative yet efficient assessment of deviation from expectations;
- **C4.1.2.1**: New multi-hypotheses correction based on a variant of Hochbert procedures able to effectively test deviations and minimize the risk of false negatives of conservative peers (type-II errors) and false positives (type-I errors);
- **C4.1.2.2**: Binary space partitioning algorithm for efficient non-conservative corrections from an arbitrary-high number of hypotheses;

R4.1.3: Guarantee the applicability to tabular data contexts with complex domains (mixtures of nominal, ordinal and numeric features);
R4.2: Robust assessment of additive, multiplicative, symmetric, order-preserving and plaid coherencies;
R4.3: Robust assessment of regions with arbitrary-high levels of noisy and missing elements;
R4.4: Robust assessment in real-valued data contexts;
R4.5: Robust assessment of regions from structured data;
R4.6: Inference of global constraints;
R4.6.1: Expectations on the properties of biclusters assuring their statistical significance of biclusters;
- **C4.6.1.1**: Global statistical tests (founded on the Poisson distribution to characterized deviations on the number of occurring biclusters) to infer minimum number of rows and columns in a bicluster that guarantees statistical significance;
- **C4.6.1.2**: Extensions to guarantee an adequate balance between type-I or type-II errors (address problems of peer contributions);
- **C4.6.1.3**: Extension of the proposed global tests to guarantee their adequate applicability over non-constant models;

R4.6.2: Effective incorporation within the learning process;
- **C4.6.2**: Principles to adequately prune of the search space of biclustering tasks in the presence of global and/or local constraints;

R4.7: Integrating homogeneity and significance views for a complete assessment of biclustering models;

Table 21: Proposed contributions to assess the statistical significance of non-constant and noisy biclustering models (*Chapter* 2).

R4.2: Robust assessment of additive, multiplicative, symmetric, order-preserving and plaid coherencies;
- **C4.2.1.1**: Assessment of additive (and symmetric) models based on the examination of patterns with different amplitudes;
- **C4.2.1.2**: Assessment of multiplicative models based on the examination of patterns who share greatest common divisors;
- **C4.2.1.3**: Assessment of order-preserving models based on the allowed linear orderings (permutations);
- **C4.2.1.4**: Assessment of plaid models by recovering the original coherence in the absence of overlapping layers;
- **C4.2.2**: Theoretically inferred equations to characterize the impact of flexible coherency assumptions on the search space;
- **C4.2.3**: Depth-first searches with dynamic programming principles to avoid redundant computations;

R4.3: Robust assessment of regions with arbitrary-high levels of noisy and missing elements;

R4.3.1: Assessment of noisy biclusters;
- **C4.3.1a**: Feature-based mode (or median for numeric/ordinal data) calculus to compute pattern expectations of constant models;
- **C4.3.1b**: Extended mode calculus with removal of adjustments to assess additive, multiplicative, symmetric and plaid models;
- **C4.3.1c**: Extension to compute mode permutations to model expectations of order-preserving models;

R4.3.2: Assessment of sparse biclusters;
- **C4.3.2**: Retrieval of pattern expectations for: 1) structurally sparse data (such as network data) where elements may not have values assigned, and 2) data with arbitrary number of missings where elements may have multi-item imputations;

R4.7: Integrating homogeneity and significance views for a complete assessment of biclustering models;
- **C4.7.1**: Effective combination of the p-value given by the probability of a given bicluster to deviate from expectations affected with the p-value associated with the probability of a bicluster to have unexpectedly low levels of noise;
- **C4.7.2**: Alternative scores for combining views: 1) adjustments by quality based on deviation from expectations; and 2) adjustments based on the area to benefit low probable patterns;

Table 22: Contributions to assess the significance of real-valued biclustering models with continuous factors (*Chapter* 3).

R4.4: Robust assessment in real-valued data contexts;

R4.4.1: Robust assessment of real-valued biclusters with (possibly unknown) coherency strength;
- **C4.4.1.1**: Principles to estimate the original coherency strength and recover the underlying coherency assumption;
- **C4.4.1.2a**: Non-biased estimators of the true significance of real-valued biclusters with lower and upper bounds (bar envelope) on the expected statistical significance;
- **C4.4.1.2b**: Extended estimators to assess (real-valued) biclusters with non-constant coherencies;
- **C4.4.1.3**: Comparison of the proposed estimators with the assessment of discretized biclusters (with multi-item assignments);

R4.4.2: Assessment of additive and multiplicative biclusters with continuous adjustment factors;
- **C4.4.2a**: Estimate-driven retrieval of the allowed ranges of shifting and scaling factors;
- **C4.4.2b**: Effective assessment of additive models based on the integral of the product of slided density functions;
- **C4.4.2c**: Effective assessment of multiplicative models based on the integral of the product of size-adjusted (scaled) density functions;
- **C4.4.2d**: Interpolation principles to guarantee the efficiency of the integral calculus;

R4.1.3: Guarantee the applicability to tabular data contexts with complex domains (mixtures of nominal, ordinal and numeric features);
- **C4.1.3.1**: Three-step extension of the proposed assessments for complex domains (tests assuming dedicated distributions per feature);
- **C4.1.3.2**: Principles to model the impact of dimensionality on significance for non-identically distributed features;
- **C4.1.3.3**: Principles to approximate the properties of the search space, and thus the applied correction;
- **C4.1.3.4**: Support for biclusters with mixtures of coherency assumptions;

Table 23: Major contributions to assess the statistical significance of local descriptive models from structured data (*Chapter* 4).

R4.5: Robust assessment of the statistical significance of local descriptive models from structured data;

R4.5.1: Significance assessment of cascade models and arrangements of events;
- **C4.5.1.1**: Related work on how to infer null models, non-parametric equations and hypothesis tests to test temporal patterns;
- **C4.5.1.2a**: Statistical view on the expected support of sequential patterns by combining sequence- and itemset-wise views;
- **C4.5.1.2b**: Hypothesis testing from this view with guarantees of deviations from expectations based on new efficient procedure to compute Hochbert corrections for large outputs and to parametrically control the rate of false positives and negatives;
- **C4.5.1.3a**: Alternative statistical view on how to assess sequential patterns from probabilistic models, by testing: 1) the weight unexpectedness associated with the probability paths of a region, or 2) the region coverage and items' dependence;
- **C4.5.1.3b**: Extensibility of statistical tests on probabilistic finite automata towards complex hidden Markov models;
- **C4.5.1.4**: Shown compliance of these contributions to assess temporal patterns and, ultimately, cascades and arrangements of events;

R4.5.2: Learning guidance from statistical significance criteria;
- **C4.5.2a**: Local statistical tests to monotonically prune the search space for deterministic methods;
- **C4.5.2b**: Principles to guide the learning of Markov-based models and their decoding step;
- **C4.5.2c**: Inference of global constraints from deviations on the Poisson distribution of expected occurrences;

VI. Learning Effective Classifiers from High-Dimensional Data

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.

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 24-29 exhaustively list the proposed contributions throughout this book.

Table 24: Major contributions to learn effective associative classifiers from high-dimensional tabular data (*Chapter* 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 25: Proposed contributions on the learning of classifiers from regions with varying homogeneity (*Chapter* 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 26: 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* 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 27: Proposed contributions to learn associative classifiers from structured data (*Chapter* 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 28: Contributions for the learning of classifiers from significantly discriminative and informative regions and with minimized propensity to overfit and underfit high-dimensional data (*Chapters* 5 and 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 29: Contributions to classify a class at different time periods for predictive tasks (*Chapter* 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.

VII. Thesis Contributions and Implications

Throughout the previous five books, we addressed different challenges associated with learning in high-dimensional data contexts. As a result, the conducted research gave rise to the following contributions.

Performance Guarantees of Models Learned from High-Dimensional Data

The first part of the thesis established a solid foundation for the assessment of descriptive and classification models in high-dimensional data contexts (contributions listed in Tables 3 to 5). First, we proposed robust estimators to bound and compare the performance of classification models, showing how existing estimators can be enhanced to: adequately measure the generalization error, accommodate smoothing factors in the presence of probabilistic outputs, and minimize the problems associated with the inference of performance guarantees from unfeasible error estimates. Second, we extended this assessment towards local descriptive models, showing how error estimates can be collected to infer robust guarantees of performance and parameterized with adequate loss functions. In this context, we overviewed the properties of existing loss functions for descriptors of tabular data (biclustering models) and structured data (triclustering and cascade models), and proposed new loss functions able to provide relevant performance views from synthetic and real data. Finally, to further guarantee non-biased and complete assessments of descriptive and classification models, we proposed data generators for (labeled) tabular and structured data based on the different forms of local and global regularities commonly observed in real-world data.

Implications

These contributions open a new door for robust, unbiased and complete assessments of state-of-the-art descriptive and classification methods. They are also critical to: 1) gain an in-depth understanding of the behavior of these methods in high-dimensional data contexts, 2) unravel their strengths and weakness, and 3) (possibly) guide their improvement. The guarantees of performance inferred under the proposed methodology can be used to weight and validate the high number of implications derived from the analysis of models learned from real data, and to measure the impact of applying procedures for dimensionality reduction prior to learning.

Learning Local Descriptive Models from Tabular Data

The second part of our thesis has contributions (enumerated in Tables 6 to 15) to the learning of robust biclustering models with flexible structures, coherency and quality from (high-dimensional) tabular data. Towards this end, we first explored the synergies between biclustering and pattern mining by studying the impact of using distinct patterns (including itemsets, association rules, formal concepts and sequences), dedicated (anti-)monotonic searches, and multiple preprocessing and postprocessing procedures on the properties of biclustering models.

Second, we extended these contributions to learn biclustering models with different forms of coherency commonly observed in biomedical and social data domains, including: additive and multiplicative models (by interpreting shifts and common multiples in data); plaid models (by identifying meaningful interactions between biclusters); order-preserving models (by discovering noise-tolerant orderings of values); and the previous models in the presence of symmetries. In particular, we proposed new algorithms to learn such non-constant models and extended them to deal with non-trivial plaid effects and be able to incorporate meaningful relaxations.

Implications

The learning of flexible biclustering models opens many possibilities for the analysis of biomedical and social data. The discovery of a

flexible structures is essential to guarantee an unbiased exploration of the search space. The discovery of additive and multiplicative coherencies is critical to analyze biomedical data with structural differences on the responsiveness associated with gene expression, molecular concentrations or physiological responses. In social applications, these coherencies are useful to model social interactions with non-trivial yet coherent behavior and to group subjects with identical variation of preferences for browsing, (e-)commercing and collaborative filtering. The accommodation of symmetries are used to capture distinct regulatory mechanisms in biological domains and opposed (yet correlated) trading, tweeting and browsing activity.

The possibility to learn plaid models using meaningful relaxations is critical to detect and interpret meaningful interactions between biological processes, clinical responses and social behavior. Similarly, the proposed composition functions also allow the study of interactions between biological processes and social groups from the point of view of their regulatory/behavioral activity, enabling the categorization of the interactions and providing insights to understand complex regulation/behavior.

Furthermore, the possibility to mine (strictly or monotonically increasing) orderings has been shown to be essential to solve different real-world problems, including the analysis of omic, chemical and mutagenesis data; the discovery of (relative) preferences from collaborative filtering data; the support to planning, scheduling and recommendation tasks; among others.

Finally, and for the first time, the proposed contributions also enable the sound application of biclustering over real-world tabular data characterized by a mixture of numeric, ordinal and categoric features.

Third, we revised the previous algorithms to guarantee their robustness to noise. In this context, we extended them with: multi-item assignments in order to guarantee that they are not exposed to discretization problems; principles along the preprocessing, mining and postprocessing steps to guarantee their robustness to varying forms and amount of noise; and multi-value imputations to handle and arbitrary-high number of missings.

Implications

These contributions provide a mark on how to effectively address the item-boundaries problem of discretization procedures by learning from data elements with an arbitrary number of items assigned. Results further show that under adequate searches this option does not visibly hamper the computational complexity of the task. As such, this option seizes the benefits of discretizing data without apparent downsides, thus opening critical considerations for future research.

Learning from data elements with multiple values is also valuable to deal with noise and provide risk-free imputations. Furthermore, its combination with the remaining principles to guarantee robustness appears to be essential to learn from data with an arbitrary-high number of missings (such as often found in biomedical domains) and for the analysis of real-world data with varying forms and amount of noise.

Fourth, we proposed new algorithms to guarantee the efficiency of the biclustering task for dense and high-dimensional data, including new pattern-growth searches for frequent itemset mining (based on annotated frequent pattern trees less prone to memory and time bottlenecks) and sequential pattern mining (based on efficient data projections and item-indexable properties to mine both monotonically and strictly increasing orders). We further guarantee their scalability in the presence of approximate searches and data partitioning principles from pattern mining. Finally, we also guarantee the efficiency of merging, extension and reduction procedures for postprocessing by placing anti-monotonic heuristics and pushing their computation to the mining step.

Implications

By integrating the well-studied principles to guarantee the scalability of pattern mining searches, the computationally complex task of learning flexible biclustering models from large-scale data becomes tractable. In this context, the proposed contributions remain prepared to handle the increasing size and dimensionality of real-world data.

Furthermore, the contributions for the efficient postprocessing of biclustering solutions can be transversally applied to manipulate voluminous outputs in the context of a wide-variety of research streams. As part of these contributions, some of the proposed mappings to solve postprocessing tasks, such as the possibility to merge an arbitrary-high number regions by relying on multi-support pattern mining tasks, open new avenues to guarantee the scalability of these procedures.

Fifth, we extended the proposed algorithms to learn from large-scale network data. For this aim, we revised the underlying data structures and searches, and further investigate the role of pattern-based biclustering to discover modules with non-dense yet meaningful and coherent interactions, as well as to guarantee their robustness to noisy and missing interactions. We showed the applicability towards homogeneous and heterogeneous networks with either quantitative/weighted or qualitative/labeled interactions. Finally, we generalized these principles to learn from sparse data, characterized by the presence of an arbitrary number of uninformative elements and/or missings.

Implications

These contributions provide the unprecedented opportunity to efficiently discover non-trivial (yet meaningful, coherent and significant) modules from network data. The discovery of these modules is essential to characterize, discriminate and/or predict biological functions and social activity. In particular, we expect to entail broader biological analysis to further establish relationships between modules and biological functions, as non-dense models might give clues about the organization of genes, proteins and metabolites, and support the characterization of molecular entities with yet unclear roles. The proposed contributions were also shown to be relevant to identify non-trivial communities from social networks where individual may have different degree of activity/involvement. Furthermore, they can also be applicable for the analysis of network traffic monitors, neural networks, structured financial transactions and coautorship/citation networks.
The proposed principles to tolerate missing and noisy interactions within the discovered modules can be further used to predict unknown interactions and to test the confidence of the existing interactions. Also, the plaid model can be used to explore the network structure and identify biological/social hubs based on the overlapping interactions between modules.

Finally, we proposed principles for biclustering tabular data in the presence of background knowledge. For this end, we revised the proposed pattern-growth searches in order to explore efficiency gains from succinct, (anti-)monotonic, convertible and prefix-monotone constraints and to accommodate annotations extracted from knowledge bases and literature, and motivated their relevance across biological and social data domains.

Implications

The incorporation of background knowledge within (pattern-based) biclustering through declarative constraints is essential to orient the task according to user expectations and other available sources of knowledges. In particular, these contributions provide the unprecedented possibility to remove uninformative regions and to focus the search on (non-trivial) regions of interest. As such, we expect to see a systematization of constraints with relevance for the analysis of biomedical and social data. Furthermore, the possibility to annotate rows and columns with terms from knowledge repositories, semantic sources and literature (such as biological functions from gene ontologies and well-studied networks) can valuably guide the learning.

We further showed that the described contributions can be soundly integrated and their synergies explored. For this end, we proposed the BicPAMS software with declarative, graphical and programmatic interfaces.

Learning Local Descriptive Models from Structured Data

The third part of this thesis has contributions (listed in Tables 16 to 19) to the learning of local descriptive models from (high-dimensional) structured data. First, we motivated and formalized the relevant task of learning local descriptive models from three-way time series given by cascades of modules. A new algorithm was proposed to discover flexible modules and their causal relation, and to handle arbitrary-high temporal misalignments between their supporting observations. We further extended the algorithm to guarantee its robustness to noise and to stochastic uncertainties associated with divergent paths, as well as to enhance its scalability based on dissimilarity guarantees, removal of uninformative elements and approximate searches under strict optimality guarantees.

Implications

The learning of cascade models from multivariate time series opens a new door to study regulatory responses to growth, development, drugs and disease progression in biomedical domains, as well as behavioral responses (associated with social interaction, web navigation, commercial activity, financial decisions) to specific events of interest in social domains. In this context, the proposed contributions provide the unprecedented opportunity to handle the inherent stochasticity and complexity of these responses and to model their structural and temporal kinetics.

Second, we proposed a new algorithm to learn local descriptive models given by arrangements of informative events from multi-dimensional databases and collections of records. In this context, we provided a structured view on the relevance of this task, and an adequate data mapping of complex data structures into multi-sets of events for a cohesive tackling of the target task. The proposed algorithm is able to discover sets of temporally-related events derived based on the discovery temporal patterns from multi-sets of events. Multiple principles were proposed to enhance its efficiency, to deal with the structural sparsity of multi-sets of events, to prevent that some arrangements jeopardize the learning, and to tolerate temporal misalignments and noise.

Implications

The proposed contributions open new opportunities to learn informative relations from complex data structures. This direction has been widely termed a hot topic in data mining due to the increasing availability, quality and high-dimensionality of large-scale structured data. Illustrating, corporative and administrative data from public and private sectors typically follow multi-dimensional or relational schema. Also, repositories of heterogeneous health-records, user actions or financial transactions are additional data structures not supported by the majority of existing learning algorithms. In this context, the provided principles to consistently map these databases as multi-sets of events and to adequately learn informative arrangements of events (combining temporal and cross-attribute dependencies) can be seen as a relevant mark in the machine learning community.

Finally, we proposed alternative stochastic methods to model cascades and arrangements of events from structured data. For this aim, we first provided a probabilistic view of itemset sequences by extending hidden Markov models with: new architectural components to model frequent orderings, principles for their effective initialization, revised learning schema to surpass convergence problems, and traversal procedures to efficiently decode sequential patterns from the learned lattices. As a result, we proposed a stochastic sequential pattern miner tolerant to noise (yet robust to spurious background matches) and able to model dissimilar, lengthy and non-trivial patterns. Finally, we shown that this algorithm can be adequately extended to stochastically learn local descriptive models from three-way time series and multi-sets of events, and confront its benefits and limitations against deterministic peers. For this end, we customized some of the proposed architectures and enhanced them with the unprecedented possibility to disclose time frames associated with state emissions.

Implications

The stochastic modeling of temporal patterns from structured data is relevant to provide compact representations of commonly large outputs; offer a probabilistic and noise-tolerant view of patterns; seize efficiency gains when learning from dense data; support the query-driven decoding of regions with patterns of interest; and focus the search on dissimilar and arbitrary-large regions. Stochastic models are of particular relevance for real-world data with complex stochastic phenomena and without well-defined regions. The proposed contributions enable the stochastic stochastic learning of descriptive models from structured data with non-fixed multivariate order and high-dimensionality. In this context, we expect to witness the extension of these contributions (such as new architectural components more conducive to the learning of certain temporal patterns), as well as their generalization for alternative stochastic learning methods, including neural networks, stochastic grammars and dynamic Bayesian networks.

Significance Guarantees of Local Descriptive Models

The fourth part of the thesis has contributions (enumerated in Tables 20 to 23) to assess and guarantee the statistical significance of regions modeled from (high-dimensional) data. For this aim, we first proposed statistical tests to robustly assess the statistical significance biclusters and enhanced the proposed biclustering algorithms accordingly. In this context, we revised non-conservative corrections (with principles to efficiently guarantee a minimized risk of accepting non-significant biclusters and rejecting significant biclusters) to guarantee their deviation from expectations. Complementarily, we shown how either local or global expectations on the size of biclusters can be inferred to guide the biclustering task.

Second, we extended the proposed statistical assessment towards noisy biclusters with flexible coherency. For this aim, new statistical views were provided to test additive and multiplicative models (based on allowed shifts and greatest common divisors), plaid models (based on the removal of plaid effects), and order-preserving models (based on the allowed permutations). The properties of search space size for these coherency assumptions were revised, and dynamic programming principles were considered to avoid redundant computations. For the assessment of noisy biclusters, we provided new principles to retrieve the underlying pattern expectations, and further extended the statistical framework to integrate homogeneity and significance views by combining the probability of a given bicluster to deviate from expectations with the probability to have unexpectedly low levels of noise.

Third, we extended the proposed statistical assessment towards real-valued biclusters with continuous factors and (possibly unknown) coherency strength. In this context, we proposed non-biased estimators of their true significance with lower and upper bounds. Additionally, new statistical tests to assess biclusters with continuous

shifts and scales were provided based on the integral of the product of slided and scaled density functions. We further guarantee the extensibility of the statistical tests in this book towards tabular data with non-identically distributed features.

Implications

Assessing the statistical significance of biclustering models is critical to filter, validate or weight the increasing number of implications in literature derived from the analysis of local regions from biomedical and social data. Furthermore, it is essential to evaluate and compare state-of-the-art biclustering and pattern mining algorithms with regards to the significance of their solutions, as well as to guide the biclustering tasks, promoting the significance of their outputs and efficiency of the searches.

The provided systemic analysis on how the properties of a given bicluster (support, length, pattern, coherency, quality) and the input data (size, dimensionality, regularities) affect statistical significance, can be further used to shape upcoming biclustering algorithms.

As largely motivated, guaranteeing the significance of biclusters is further relevant to minimize the propensity of a given learning function to over/underfit the observed data, ensuring the adequacy of the capacity term and thus its ability to generalize.

Finally, we provided methods to assess the statistical significance of local descriptive models from structured data. In this context, we provided statistical views to test the deviation of the support of a sequential pattern against expectations derived from null data, as well as alternative tests on probabilistic models (compliant with hidden Markov models) based on the weight unexpectedness and coverage of a sequential pattern. These assessments were extended towards temporal patterns, cascades and arrangements of events, and incorporated within the previously proposed deterministic and stochastic algorithms to guide the learning.

Implications

The proposed contributions are essential to assess and guarantee that regions in structured data contexts are not informative by chance, a prominent problem due to the typical high-dimensionality of structured data given by multivariate time series, multi-sets of events and multi-dimensional data.

Learning Effective Classifiers from Local Descriptive Models

The fifth part of this thesis has contributions (listed in Tables 24 to 29) to the learning of effective (associative) classifiers from high-dimensional data. Towards this end, we first proposed new associative classifiers with principles to guarantee the adequacy of their discovery, training and testing functions. To guarantee the discovery of relevant regions we proposed: a new weighted notion of support to adequately assess the discriminative power of noisy regions, an adequate exploration of the data space, and the inclusion of regions able to discriminate groups of classes by composing rules with disjunctions of labels. New integrative scores (able to combine the discriminative power, size, coherency and quality of a region) were proposed along with adequate composition criteria to guarantee the adequacy of training functions. Finally, for an effective testing of new observations, we proposed new relaxations on the matching criterion to prevent the scarcity of matched regions and a new calculus of class strength.

Second, to guarantee the adequate inference of decisions in data domains characterized by regions with varying homogeneity and quality, we extended the previous classifiers to guarantee the recovery of with varying coherency strength, coherency assumption and noise. A new penalization schema was proposed for non-constant regions based on their degree of flexibility to prevent that they jeopardize the learning, as well as new matching criteria to test observations against non-constant regions.

Implications

The proposed associative classifiers are essential to guarantee an adequate learning from high-dimensional data domains characterized by the presence of regions of interest (possibly) well-approximated by discriminative biclusters. In this context, they are particularly relevant to learn biological and clinical markers, classify user behavior, evaluate (web) contents, and support trading/administrative/commercial decisions. They are further proposed as viable alternatives to surpass the largely motivated problems associated with feature selection, dimensionality reduction and/or sparse priors.

The proposed contributions provide an unprecedented opportunity to address 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. Furthermore, contrasting with the focus on constant regions given by discriminative patterns of existing associative classifiers, the proposed associative classifiers are able to retrieve the commonly observed non-constant (yet meaningful and coherent) regions underlying biomedical and social data.

These contributions further enable the systematic analysis on how the varying coherency, quality, size and discriminative power of the underlying regions of a high-dimensional data space affect the performance of classifiers. This understanding is essential to revise the behavior of state-of-the-art classifiers and place considerations on the design of new classifiers.

Third, we further extended the previous associative classifiers to guarantee their ability to: learn from sparse data by soundly removing true missings and uninformative elements and adequately interpreting false missings; learn from data with global and local regularities by robustly combining the (probabilistic) outputs of alternative classifiers; learn from stochastic descriptors of tabular data where membership vectors can be used to promote an adequate space coverage and shape the scoring criteria; and learn from data in the presence of background knowledge by effectively incorporating annotations and constraints with nice properties.

Implications

These contributions provide the unprecedented opportunity to classify sparse data, network data and data with missing elements (associated with monitoring holes, default expectations and errors). Also, they introduce the possibility to remove non-interesting regions to guide the learning and explore efficiency gains. In biological data domains, uninformative elements are typically associated with non-differential regulation of genes or concentration of molecular entities. For other domains, uninformative elements may correspond to: entries with low-counts from text-based data, inconclusive ratings in collaborative filtering data, unprofitable decisions from trading data, or healthy evaluations from medical data.

Finally, the possibility to incorporate a large variety of constraints and knowledge-driven annotations provides the opportunity to flexibly guide the learning of (associative) classifiers.

Fourth, we proposed new associative classifiers able to effectively learn from structured data. For this aim, the previously proposed deterministic algorithms to discover regions from structured data were extended with discriminative criteria. From these regions, rules are inferred, ranked according to a new integrative score and composed within a navigable tree structure. During the testing phase, we proposed new matching criteria sensitive to both structural and temporal misalignments, and revised the class strength calculus accordingly. Principles for selective/decaying memory can be easily incorporated within these classifiers. Complementarily, stochastic classifiers were proposed by testing the likelihood of new observations to be described by class-conditional probabilistic models of a data partition. Finally, we show the applicability of these classifiers to learn from varying data structures, including multivariate time series, itemset sequences, multi-sets of events and multi-dimensional databases.

Implications

The proposed contributions are key 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.

Contrasting with existing classifiers, the proposed associative classifiers provide the possibility to learn from multi-sets of events, handling arbitrary levels of sparsity between events and integrating distinct event types (heterogeneous attributes). Similarly, the proposed classifiers are also critical to learn from regions of three-way time series given by discriminative cascades.

Finally, the provided comparison between deterministic and stochastic learners offer key principles for an adequate selection and parameterization of classifiers according to the end goal and the properties of the input data.

Fifth, we proposed a structured view on how to assess and shape the guarantees of statistical significance of classifiers learned from high-dimensional data. For this purpose, we first discussed the benefits from learning from statistically significant regions and extended the previously proposed statistical views of regions to also assess the significance of their discriminative power. This statistical view was used to revise the behavior of associative classifiers, decision trees and random forests, thus minimizing their propensity to underfit high-dimensional data.

We further extended these statistical views to assess the impact that training and testing functions have on the risk towards false positive and negative decisions. To turn these views transparent to the user, we proposed the annotation of rules and classification decisions with an indicative score of their guarantees of statistical significance. Moreover, in order to avoid the blind optimization of the behavior of classifiers according to this criteria, we revised their learning according to both accuracy (average error) and significance (variability of error) views.

Implications

Guaranteeing that classification decisions are inferred from statistically significant models is of heightened importance to learn biological and clinical markers and to support computer-aided decisions associated with medical/trading/marketing/administrative initiatives with either high impact on daily lives or high costs.

These contributions are also vital to guarantee the adequacy of learning from high-dimensional data with a (possibly) limited number of observations. In particular, they are decisive to surpass the: 1) overfitting risk of global classifiers (by guaranteeing that uninformative regions are discarded); 2) underfitting risk of classifiers reliant on procedures for dimensionality reduction (by preventing the loss of relevant regions and the inclusion of new forms of bias); and 3) underfitting risk of local classifiers (by guaranteeing that decision rules are inferred from significantly informative and discriminative regions).

The provided results stress the importance of embracing sound statistical views to assess the propensity of a given classifier to make false positive and false negative decisions. As such, new directions for future work become necessary: the extension of the provided experimental analysis; the exploration of the provided statistical principles to further guide the learning of local classifiers; and the generalization of the proposed contributions towards alternative classification models, such as support vector machines, neural networks and Bayesian classifiers.

Sixth and finally, we tackled the multi-period classification task by proposing new algorithms able to: 1) learn sequences of classes with intricate stochastic dependencies, and 2) embed (single-label) classifiers to adequately learn from tabular and structured data. In this context, we motivated and formalized the task, studied meaningful metrics (sensitive to temporal misalignments and error accumulation on the estimated sequences) to robustly assess multi-period classifiers, and proposed new algorithms reliant on clustering for an adequate reduction and recovery of the space of sequences. Variants of the clustering-based algorithms were proposed based on the segmentation of the sequence of classes (based on their local periodicities) and on the use of sliding-windows to minimize the risk towards false positive and false negative decisions.

Implications

Multi-period classification is essential to answer a wide-set of real-world problems. The prediction of the evolving state of living, geophysical, economic and societal systems (referred as one of the ten most critical data mining challenges for this decade [33]) is necessary to: anticipate epidemics or environmental changes; assess key changes in human health (from clinical, psychophysiological and biological perspectives) and human behavior (based social data); support personalized decisions (prognostics); characterize disease progression; and support administrative decisions in both private industries and public sectors and planning tasks.

The proposed contributions to model the stochastic dependencies between the periods under classification and embed existing learning functions opens up new possibilities for the application of existing classifiers for predictive tasks from varying data structures.

Bibliography

[1] R. Rathipriya K. Thangavel J. Bagyamani. Binary particle swarm optimization based biclustering of web usage data. *CoRR*, abs/1108.0748, 2011.

[2] Arindam Banerjee, Inderjit Dhillon, Joydeep Ghosh, Srujana Merugu, and Dharmendra S Modha. A generalized maximum entropy approach to bregman co-clustering and matrix approximation. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 509–514. ACM, 2004.

[3] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer, 2011.

[4] Stanislav Busygin, Nikita Boyko, Panos M Pardalos, Michael Bewernitz, and Georges Ghacibeh. Biclustering eeg data from epileptic patients treated with vagus nerve stimulation. In *Data mining, systems analysis and optimization in biomedicine*, volume 953, pages 220–231. AIP Publishing, 2007.

[5] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[6] P.A.D. de Castro, F.O. de Franga, H.M. Ferreira, and F.J. Von Zuben. Applying biclustering to perform collaborative filtering. In *Intelligent Systems Design and Applications*, pages 421–426, Oct 2007.

[7] Chris Ding, Ya Zhang, Tao Li, and Stephen R. Holbrook. Biclustering protein complex interactions with a biclique finding algorithm. In *ICDM*, pages 178–187, Washington, DC, USA, 2006. IEEE Computer Society.

[8] Pedro Domingos. A unified bias-variance decomposition and its applications. In *IC on Machine Learning*, pages 231–238. Morgan Kaufmann, 2000.

[9] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2790–2797, June 2009.

[10] Gang Fang, Majda Haznadar, Wen Wang, Haoyu Yu, Michael Steinbach, Timothy R. Church, William S. Oetting, Brian Van Ness, and Vipin Kumar. High-Order SNP Combinations Associated with Complex Diseases: Efficient Discovery, Statistical Power and Functional Interactions. *Plos One*, 7, 2012.

[11] Mário AT Figueiredo and Anil K Jain. Bayesian learning of sparse classifiers. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–35. IEEE, 2001.

[12] Mário AT Figueiredo, Anil K Jain, and Martin H Law. A feature selection wrapper for mixtures. In *Pattern Recognition and Image Analysis*, pages 229–237. Springer, 2003.

[13] Mário AT Figueiredo and Robert D Nowak. Wavelet-based image estimation: an empirical bayes approach using jeffrey's noninformative prior. *Image Processing, IEEE Transactions on*, 10(9):1322–1331, 2001.

[14] Dmitry Gnatyshak, DmitryI. Ignatov, Alexander Semenov, and Jonas Poelmans. Gaining insight in social networks with biclustering and triclustering. In *Perspectives in Business Informatics Research*, volume 128 of *Lecture Notes in Business Info. Processing*, pages 162–171. Springer Berlin Heidelberg, 2012.

[15] R. Henriques and C. Antunes. Learning predictive models from integrated healthcare data: Extending pattern-based and generative models to capture temporal and cross-attribute dependencies. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, pages 2562–2569, Jan 2014.

[16] R. Henriques and S. Madeira. Biclustering with flexible plaid models to unravel interactions between biological processes. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 2015.

[17] Rui Henriques. *Mining emerging patterns to construct accurate and efficient classifiers*. PhD thesis, Instituto Superior Tecnico, Universidade de Lisboa, Lisboa, 2016.

[18] Rui Henriques, Cláudia Antunes, and SaraC. Madeira. Generative modeling of repositories of health records for predictive tasks. *Data Mining and Knowledge Discovery*, 29(4):999–1032, 2015.

[19] Rui Henriques and Sara Madeira. Bicpam: Pattern-based biclustering for biomedical data analysis. *Algorithms for Molecular Biology*, 9(1):27, 2014.

[20] Rui Henriques and Sara Madeira. Bicspam: Flexible biclustering using sequential patterns. *BMC Bioinformatics*, 15:130, 2014.

[21] Rui Henriques and Sara C. Madeira. Bicnet: Efficient biclustering of biological networks to unravel non-trivial modules. In *Algorithms in Bioinformatics (WABI)*, LNCS. Springer-Verlag, 2015.

[22] Rui Henriques and SaraC. Madeira. Towards robust performance guarantees for models learned from high-dimensional data. In Aboul Ella Hassanien, Ahmad Taher Azar, Vaclav Snasael, Janusz Kacprzyk, and Jemal H. Abawajy, editors, *Big Data in Complex Systems*, volume 9 of *Studies in Big Data*, pages 71–104. Springer International Publishing, 2015.

[23] Qinghua Huang. A biclustering technique for mining trading rules in stock markets. In Dehuai Zeng, editor, *Applied Informatics and Communication*, volume 224 of *Communications in Computer and Information Science*, pages 16–24. Springer Berlin Heidelberg, 2011.

[24] L. Lazzeroni and A. Owen. Plaid models for gene expression data. *Statistica Sinica*, 12:61–86, 2002.

[25] Jinze Liu and Wei Wang. Op-cluster: Clustering by tendency in high dimensional space. In *ICDM*, pages 187–, Washington, DC, USA, 2003. IEEE Computer Society.

[26] Sara C. Madeira and Arlindo L. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 1(1):24–45, January 2004.

[27] Yoshifumi Okada, Kosaku Okubo, Paul Horton, and Wataru Fujibuchi. Exhaustive search method of gene expression modules and its application to human tissue data. *IAENG IJ of Comp. Science*, 34(1):119–126, 2007.

[28] Akdes Serin and Martin Vingron. Debi: Discovering differentially expressed biclusters using a frequent itemset approach. *Algorithms for Molecular Biology*, 6:1–12, 2011.

[29] Miranda van Uitert, Wouter Meuleman, and Lodewyk Wessels. Biclustering sparse binary genomic data. *Journal of Computational Biology*, 15(10):1329–1345, 2008.

[30] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[31] Shu Wang, Robin R Gutell, and Daniel P Miranker. Biclustering as a method for rna local multiple sequence alignment. *Bioinformatics*, 23(24):3289–3296, 2007.

[32] Eric P Xing, Michael I Jordan, Richard M Karp, et al. Feature selection for high-dimensional genomic microarray data. In *ICML*, volume 1, pages 601–608. Citeseer, 2001.

[33] Qiang Yang and Xindong Wu. 10 challenging problems in data mining research. *International Journal of Information Technology and Decision Making*, 5(4):597–604, 2006.