FleBiC: Learning classifiers from high-dimensional biomedical data using discriminative biclusters with non-constant patterns

Rui Henriques[a], Sara C. Madeira[b]

[a]*INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal*
[b]*LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal*
Contact: rmch@tecnico.ulisboa.pt, sacmadeira@fc.ul.pt

Abstract

The discovery of discriminative patterns from high-dimensional data offers the possibility to learn from informative subspaces and pattern-centric features, paving the way to associative classifiers. Despite the success achieved by associative classifiers, such as random forests or XGBoost, they generally neglect discriminative subspaces with non-constant coherencies. Research on biclustering has for two decades highlighted the role of non-constant patterns in biomedical domains, including additive and order-preserving patterns. Still, their relevance for classification remains unexplored.

This work assesses the impact of discriminative patterns with varying coherence and quality on associative classification. A novel classifier, FleBiC, is proposed as a result. FleBiC extends pattern-based biclustering with principles to match observations against non-constant and noise-tolerant patterns, address generalization difficulties, minimize scarcity of matches, support class disjunctions, and offer statistical guarantees. Results on biological and clinical data highlight the role of non-constant patterns, specially order-preserving patterns, for improving the performance of state-of-the-art classifiers.

Keywords: associative classification, discriminative paterns, biclustering, non-constant patterns, biomedical data, high-dimensional data

1. Introduction

Learning accurate classifiers from high-dimensional data is increasingly relevant across biomedical domains [1]. High-throughput technologies currently enable the large-scale profiling of biological entities per sample (observation), including the expression of thousands of genes or the concentration of hundreds of molecular compounds. Integrative healthcare systems give rise to hundreds of physiological records per patient (observation). In these data contexts, observations are often labeled according to their phenotypes,

such as morphology, pathology, clinical assessments, applied treatment, response to stimuli, or development stage. These labeled observations can then be used to characterize phenotypes and support real-world decisions. However, learning biomedical markers and classifiers is challenged due to the: 1) complexity of the underlying biological and physiological patterns [2], and 2) high-dimensionality of biomedical data [3].

First, biomedical data is heterogeneous and typically characterized by the presence of both constant and non-constant patterns. *Constant patterns* – value expectations on a subset of variables (e.g. genes with specific levels of expression) verified on a subset of observations (e.g. individuals with a given phenotype) – are the paradigmatic case. In contrast with constant patterns, *non-constant patterns* allow value expectations to vary across observations as long as these variations are coherently explained by certain factors. Illustrating, two individuals may have a subset of genes with different expression profiles, yet these differences can be coherently explained by an additive factor (genes coherently activated or repressed yet with lower levels of expression for one of the individuals), or an order-preserving factor (identical ranks of expression irrespectively of the absolute values). The presence of such factors gives rise to non-constant patterns, including additive, multiplicative and order-preserving patterns. These factors are generally driven by:

1. varying profile of individuals and their unique biophysiological responsiveness to certain conditions;
2. varying morphology and other observable traits of the clinical conditions under study;
3. assessments along different stages of disease progression or variations on the treatment protocol; and
4. experimental and preprocessing biases per observation.

In this context, the traditional focus on constant patterns easily neglects other less-trivial yet relevant patterns that could aid the learning.

Furthermore, the inherent high-dimensionality of most biomedical datasets increases the susceptibility of classifiers to overfitting and underfitting [4]. Although feature selection, data space transformations and sparse kernels have been largely proposed to reduce dimensionality [5, 6, 7], these procedures generally disregard the presence of discriminative patterns. As such, these approaches often neglect subspaces of potential interest, thus unaddressing underfitting risks, and are unable to flexibly discard uninformative subspaces, thus unaddressing overfitting risks [8, 9]. Conceptually, these risks could be minimized by focusing the learning on all subspaces of major interest, which can be potentially given by discriminative patterns [10]. In fact, to address the curse of high-dimensionality, an increasing number of classifiers able to learn

from discriminative patterns, generally referred as associative classifiers, have been proposed [11, 12, 13]. Notable cases of associative classifiers include: 1) tree-based classifiers such as decision trees or random forests (where a pattern corresponds to a path from root to leaf with the value expectations on a subset of variables); 2) rule-based classifiers (where a pattern corresponds to the antecedent of a class-discriminative association rule); and 3) any classic classifier learned from features extracted using discriminative pattern mining and biclustering [14]. Despite their relevance, associative classifiers are generally only able to learn from constant subspaces [8, 11, 13].

These observations lead us to the following *research questions*: Can the performance of (associative) classifiers be improved in the presence of non-constant patterns? This question leads us to a second research question: How does the behavior of (associative) classifiers vary with the underlying pattern coherence and quality? In other words, to which extent do non-constant and noise-tolerant patterns impact performance?

To answer these questions, we propose a new associative classifier, FleBiC (Flexible Biclustering-based Classifier), able to learn from non-constant subspaces with parameterizable coherence and quality using principles from the biclustering field of research. FleBiC relies on three major steps: 1) discovery and composition of discriminative patterns with diverse coherence and quality (*discovery*); 2) scoring of the identified patterns (*training*); and 3) effective classification of new observations against the scored patterns (*testing*). In this context, FleBiC combines the following methodological contributions:

C1. *discovery*: extension of state-of-the-art pattern-based biclustering algorithms [2] to discover discriminative patterns with parameterizable coherence (constant, additive, multiplicative, order-preserving and plaid assumptions) and quality. As part of this contribution, we propose:

- the use of integrative biclustering searches to exhaustively discover patterns with varying coherence and quality, thus minimizing the known problem of scarce matches faced by associative classifiers [10];
- a new notion of support and confidence (weighted by the amount of tolerated noise) to better assess the discriminative power of a pattern;
- associations between patterns and disjunctions of labels in order to identify patterns able to discriminate more than one class;

C2. *training*: effective learning functions. Specifically, we propose:

- a new score combining the (weighted) support, length, quality (deviations from pattern expectations) and statistical significance of a pattern. This tackles the problem of small patterns being prioritized due to an overemphasis on their confidence;
- a penalization for non-constant coherencies to prevent the exclusion of simpler patterns (such as constant patterns) from decisions.

C3. *testing*: decisions sensitive to non-constant and noise-tolerant patterns. In particular, we propose:

- an integrated score to compute the probability of a new observation being labeled with a given class using multiple interestingness criteria (support, quality, discriminative power, significance, size, coherence);
- new matching criteria for non-constant patterns;
- relaxations to guarantee an adequate number of matches in order to minimize under/overfitting risks.

These contributions are integrated within FleBiC and their relevance are empirically shown in synthetic and real data. The gathered results show the superiority of FleBiC against peer classifiers and its unique ability to unravel new discriminative patterns in biological and clinical data domains. The results strongly support the relevance of considering non-constant patterns to improve the performance of associative classifiers, opening new considerations for research in the classification field.

The paper is structured as follows. *Section 2* provides the motivation and background for the targeted task. *Section 3* surveys contributions and limitations from related work. *Section 4* describes the solution space by proposing FleBiC. *Section 5* gathers experimental results that provide initial evidence of the superiority and utility of FleBiC. Finally, concluding remarks and implications are synthesized.

2. Background

Def. 2.1 *Given a set m variables $\mathbf{Y}=\{y_1,..,y_m\}$, a* ***dataset*** *is a set of n observations $\mathbf{X}=\{\mathbf{x}_1,..,\mathbf{x}_n\}$ where each observation $\mathbf{x}_i$ is described by a set of numeric and/or categoric values, $a_{ij} \in \mathcal{Y}_j$ where $\mathcal{Y}_j$ is the domain of $\mathbf{y}_j$ variable. Given a set of classes, $\mathcal{C}$, a* ***labeled dataset*** *is the set of pairs*

$$\{(\mathbf{x}_i, c_i) \mid i = 1..n, c_i \in \mathcal{C}\}.$$

Def. 2.2 *Given a dataset, a* ***bicluster****, $\mathbf{B}=(\mathbf{I},\mathbf{J})$, is a* subspace *defined by a subset of observations, $\mathbf{I} \subseteq \mathbf{X}$, with a pattern $\varphi_{\mathbf{J}}$ on a subset of variables, $\mathbf{J} \subseteq \mathbf{Y}$, satisfying certain criteria of interest.*

Biomedical data are characterized by the presence of coherent subspaces generally associated with putative regulatory or physiological mechanisms [15]. Each subspace (bicluster) is a subset of samples/individuals, $\mathbf{I}$, with a coherent pattern $\varphi_{\mathbf{J}}$ on a subset of biological entities/clinical tests, $\mathbf{J}$. Table 1 lists meaningful subspaces found across biomedical data domains.

Domain	*Illustrative biclusters with relevance for learning tasks*
clinical [16, 17]	Patients with correlated clinical profile (symptoms, diagnoses, prescriptions).
proteomics/metabolomics [10]	Molecular compounds with correlated concentrations on subsets of samples.
gene expression [15, 18]	Co-expressed genes involved in specific functional processes and pathways.
structural variations [19]	Correlated groups of mutations and copy number variations.
growth phenotype data [20]	Strain changes generating similar patterns of essentiality and dispensability.
biological networks [21, 22]	Modules of genes, proteins or metabolites with coherent interactions.
physiological signals [17]	Sliding features with coherent values across case or stimuli-elicited responses.
genome-wide [23]	Conserved alignments, factor binding sites and insertion mutagenesis.
other	Local patterns in translational [22], chemical [24] and nutritional data [25].

Table 1: Relevance of (discriminative) biclusters across biomedical data domains.

Def. 2.3 *Given a labeled dataset, the* ***biclustering*** *task aims to identify a set of biclusters* $\mathcal{B}=\{\mathbf{B}_1,..,\mathbf{B}_p\}$*, where each bicluster,* $\mathbf{B}_k$*, must satisfy specific criteria of* homogeneity*, statistical* significance *and* discriminative power*.*

The ***homogeneity*** *of a bicluster specifies the allowed forms of correlation among values (pattern).*

The statistical ***significance*** *of a bicluster is the probability of its pattern to occur against expectations.*

The ***discriminative power*** *of a bicluster defines its probability to only occur on a subset of classes,* $C \subset \mathcal{C}$*.*

The biclustering task is the generalization of the classic pattern mining task, originally proposed to learn from transactional data, to: 1) enable the discovery of patterns in real-valued, symbolic or mixed data (non-iid variables); 2) mine patterns with non-constant coherence and parameterizable tolerance to noise; and 3) provide guarantees of statistical significance.

Biclustering algorithms place specific criteria of interest to guide the discovery of patterns (Def. 2.3). These criteria determine the *structure*, *coherence* and *quality* of a biclustering solution. The **structure** is defined by the number, size, shape and positioning of biclusters. The **coherence** of a bicluster is defined by the observed correlation of values (coherence assumption) and the allowed deviation from perfect correlation (coherence strength). The **quality** of a bicluster is defined by the type and amount of tolerated noise.

Figure 1 presents a flexible biclustering structure (biclusters with arbitrary shape and positioning) and different coherence assumptions (Def. 2.4 to 2.6).

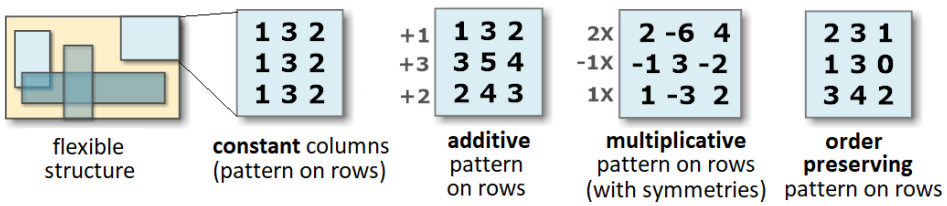

Figure 1: Flexible structure of biclusters and illustrative patterns with varying coherence.

Def. 2.4 *Given a real-valued dataset, let the elements in a bicluster* $a_{ij} \in (\mathbf{I}, \mathbf{J})$ *have* ***coherence*** *across observations given by* $a_{ij}{=}k_j{+}\gamma_i{+}\eta_{ij}$*, where* k_j *is the expected value for variable* y_j*,* γ_i *is the adjustment for observation* $\mathbf{x}_i$*, and* η_{ij} *is the noise factor. A bicluster satisfying a specific* coherence strength*,* $\delta \in \mathbb{R}^+$*, has values described by* $a_{ij}{=}k_j{+}\gamma_i{+}\eta_{ij}$ *and* $\eta_{ij}{\in}[-\delta/2, \delta/2]$*.*

Def. 2.5 *Given a bicluster* $(\mathbf{I}, \mathbf{J})$ *with coherence in accordance with Def. 2.4, the* γ_i *factors define the* coherence assumption*:* ***constant*** *columns (pattern on rows) when* $\gamma_i{=}0$*;* ***additive*** *pattern on rows when* $\gamma_i{\neq}0$*; and* ***multiplicative*** *pattern on rows if* a_{ij} *is better described by* $k_j\gamma_i + \eta_{ij}$*.*

The bicluster **pattern** $\varphi_{\mathbf{J}}$ *is the set of expected values in the absence of adjustments and noise* $\{k_j \mid y_j \in \mathbf{J}\}$*, where* k_j *is defined according to Def. 2.4.*

Consider the illustrative additive bicluster $\mathbf{B}$ in $\mathbb{N}_0^+$ from Figure 1. Assuming $\mathbf{B}{=}(\{\mathbf{x}_1, \mathbf{x}_4, \mathbf{x}_5\}, \{\mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_8\})$, this bicluster is described by $a_{ij}{=}k_j + \gamma_i$ with the (non-constant) *pattern* $\varphi{=}\{k_1{=}0, k_2{=}2, k_3{=}1\}$, supported by three observations with additive factors $\gamma_1{=}1$, $\gamma_4{=}3$ and $\gamma_5{=}2$. As our purpose is learning from labeled observations, we seek discriminative biclusters with patterns on observations (rows) and, understandably, discard biclusters with transposed coherencies given by patterns on variables (columns) [26].

Def. 2.6 *A bicluster* $(\mathbf{I}, \mathbf{J})$ *is said to satisfy an* ***order-preserving*** *coherence assumption when the values of variables* $\mathbf{J}$ *induce the same linear ordering for each observation* $\mathbf{x}_i$ *in* $\mathbf{I}$*. An* order-preserving pattern $\varphi_{\mathbf{J}}$ *thus specifies a permutation of variables,* $\pi(\mathbf{J})$*.*

Consider the illustrative order-preserving bicluster $\mathbf{B}$ in $\mathbb{N}_0^+$ from Figure 1. All observations $\mathbf{x}_i$ satisfy the permutation $\pi(\mathbf{J}){=}y_3 \leq y_1 \leq y_2$.

In accordance with previous definitions, *constant* and *non-constant* biclusters (informally referred as **constant** and **non-constant patterns**) are respectively characterized by $\forall_{\mathbf{x}_i \in \mathbf{I}}\ \gamma_i{=}0$ and $\exists_{\mathbf{x}_i \in \mathbf{I}}\ \gamma_i{\neq}0$.

Table 2 motivates the relevance of discovering biclusters with non-constant patterns, highlighting their role when learning from biomedical data.

The discovery of discriminative biomedical patterns can be pursued to characterize and differentiate phenotypes (*descriptive* setting) and support clinical decisions (*predictive* setting). In fact, one can see a discriminative pattern as a (structured) biomedical marker.

Def. 2.7 *Given a labeled dataset, an* ***associative model*** *is a composition of* p association rules*,* $f(R_1, .., R_p)$*, where* $R_i : \varphi_{\mathbf{J}_i} \Rightarrow^s C_i$ *maps a discriminative pattern* $\varphi_{\mathbf{J}_i}$ *(rule's antecedent) into a set of labels* $C \subset \mathcal{C}$ *(rule's consequent) with a given score* s*. The* composition *function* f *guarantees the effective traversal of rules according to their properties.*

Coherence	*Illustrative biclusters with non-constant patterns*
Additive and Multiplicative	Additive and multiplicative patterns accommodate shifting and scaling factors on the values across observations (Figure 1). Illustrating, two genes may be regulated in the same subset of conditions (variables) but show different expression levels explained by a shifting or scaling factor associated with their distinct responsiveness, or the bias introduced by the applied measurement and preprocessing [15]. These factors are also critical to account for subject-specific differences associated with default psychophysiological behavior and monitoring [17].
Order Preserving	Order-preserving biclusters were originally proposed to find co-expressed genes within a temporal progression (such as stages of a disease or drug response) [27]. They have been also pursued in biological data contexts where molecular concentrations of proteins and metabolites coherently vary across samples [28]. This coherence has been further applied to find sets of nodes in biological networks with an order-preserving degree of influence across another set of nodes [21, 24]. In clinical data contexts, order-preserving are essential to deal with the impact of analyzing individuals at varying stages within the progression of a disease [10]. Order-preserving biclusters can emulate constant, additive and multiplicative coherencies, leading to more inclusive solutions associated with larger modules with less susceptibility to noise.

Table 2: Relevance of non-constant biclusters when learning from biomedical data.

Moving from descriptive settings (Def. 2.7) to predictive settings (Def. 2.8) becomes a matter of defining effective training and testing functions.

Def. 2.8 *Given a labeled dataset with observations in $\mathcal{X}$, a* ***classification model*** *is a mapping function between observations and classes, $M : \mathcal{X} \rightarrow \mathcal{C}$, for labeling (unlabeled) observations.*

In particular, given an associative model (Def. 2.7), an ***associative classification model*** *relies on the f composition of discriminative patterns to learn one model to label new observations, $M(\mathbf{x}_{new}|f(R_1, .., R_p))$.*

Rule-based classifiers such as CMAR, decision trees, and random forests are notable cases of associative classifiers where the composition function places discriminative patterns within tree structures [29, 30, 12].

In this context, the learning of associative classification models from relevant patterns is driven by three major requirements:

- effective discovery of coherent and discriminative patterns;
- adequate scoring and composition of association rules based on these patterns (*training* function);
- effective matching schema to test a new observation against the scored patterns (*testing* function).

As illustrated in Figure 2, learning associative classifiers is a well-established way of dealing with high-dimensional data [9, 8]. State-of-the-art associative classifiers, such as XGBoost [31], rely on discriminative patterns, offering

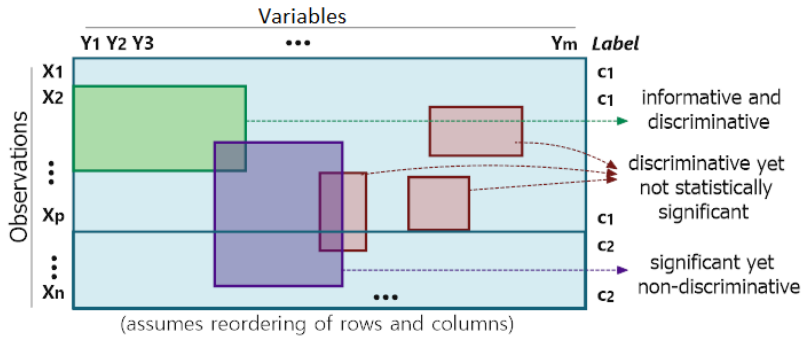

Figure 2: Role of discriminative patterns to learn classifiers in high-dimensional spaces.

a way of selecting all (and only) relevant subspaces, thus minimizing overfitting and underfitting risks. In contrast, non-associative classifiers typically place alternative principles to reduce dimensionality. *Feature selection*, ($\mathbf{I}{=}\mathbf{X}, \mathbf{J}{\subseteq}\mathbf{Y}$), generally focus on a single subspace and often neglects subspaces of potential interest, whose relevance might be only observed on a subset of observations. Classifiers based on mixtures, structured generative models and support vectors are able to learn from high-dimensional data by relying on *sparse priors* and *hyperdimensional transformations* [1, 7]. Still, these options are insufficient to find a flexible structure of arbitrarily positioned subspaces, being prone to: *1)* include non-relevant subspaces (increasing the overfitting risk), and *2)* exclude relevant subspaces (increasing the underfitting risk) [9, 1].

Given a (high-dimensional) dataset, the task targeted in this work can be formulated as learning effective associative classifiers from discriminative patterns with parameterizable coherence and quality, and assess the impact of non-constant coherence assumptions on the classification performance.

3. Related Work

Associative classification. The discovery of relevant patterns in labeled data has been mainly driven by research on information theory [32], discriminative pattern mining [12], discriminative matrix factorization [33], and, more recently, discriminative biclustering [17, 34]. Given a set of patterns of interest, different composition functions have been considered to learn an associative model, ranging from ordered sets of weighted association rules to more structured models. Examples include the integration of these patterns within Bayesian classifiers [35], decision trees [36, 37], and support vector machines (SVMs) [38]. Carreiro et al. [17] surveyed additional ways of composing biclusters from gene expression data to improve classification. In this context, adequate scoring methods are required to weight the interestingness of each pattern, ranging from simple metrics, such as the support-and-length (an indicator of the pattern's significance) and the confidence (an indicator

of discriminative power) of each rule, towards more complex metrics based on probabilistic induction [39] and optimization criteria [40]. More recently, association rules have been extended to allow for both disjunctive patterns (patterns $\varphi_{\mathbf{J}}$ with more than one expected value per variable in $\mathbf{J}$) and disjunctions of classes (patterns discriminating more than one class) [41].

Multiple testing functions over the identified patterns have been also proposed to robustly classify new observations [17, 11, 12, 37]. Illustrating, given a new observation $\mathbf{x}$, CMAR [29] retrieves association rules with exact pattern matching and computes the classes' strength $P(c \in \mathcal{C}|\mathbf{x})$ using a weighted χ^2 calculus. However, the exact matching criterion is restrictive in many data contexts since it can lead to a small (possibly empty) set of matched rules, neglecting relevant patterns due to the presence of noise. To tackle this problem, relaxations on the matching functions (allowing, for instance, the presence of shifts or the matching of a subset of overall values), as well as penalizations (affected by the extent and differences on the matching values) have been proposed [42]. Lazy classifiers that retrieve classification rules once a new observation is given are also able to address this challenge [43, 17]. In the presence of massive data, CARs-Lands [11] matches a test instance against association rules from nearest training data chunks.

Discovery of discriminative patterns. The search for discriminative patterns has received a wide-attention in literature, with multiple works providing categorizations on how to use *discriminative pattern mining* to learn associative classifiers [12, 8, 13]. Bringmann et. al [12] categorize these searches along two axes: whether they discover a pre-computed set of patterns apriori or iteratively discover new (or extend existing) patterns, and whether the search is guided or not by the properties of the target model. On the first axis, earlier studies focus on mining all constant patterns per class at a time. From the found set of patterns, many metrics have been proposed to fix adequate class-conditional support levels and to assess the correlation strength ϕ between a pattern and a class, as well as to relate both these views within a single score [12]. Illustrative associative classifiers with alternative scoring schema include CBA [44] (ϕ=confidence), classifiers based on emerging patterns [45] (ϕ=growth), CMAR [29] (ϕ=χ^2), CPAR [46] (ϕ=foil gain) and RCBT [47] (ϕ based on top-k covering rule groups). Even in the presence of constraints and condensed representations to deliver compact sets of distinct patterns, these methods are computationally expensive for large datasets or low supports. Contrasting, branch-and-bound or iterative-deepening searches avoid the generation of the complete pattern set [13], including decision trees [48], Harmony [49], DDPMine [50], MbT [51]. On the second axis, model-dependent approaches rely on the properties of the classifier to affect the discovery/selection of patterns [12]. Recently, classifiers based on ensembles

of pattern-sets, discriminative patterns found from engineered features [31], and sampling procedures [52] were shown to achieve distinctive performance in several real-world classification problems [53].

Despite its relevance, discriminative pattern mining has major limitations:

L1. inability to discover non-constant patterns, thus preventing the retrieval of non-trivial yet meaningful patterns commonly present in biomedical data domains (see Table 2);

L2. inability to discover (real-valued) patterns robust to noise, thus being susceptible to the inherent stochasticity of biological and physiological systems and biases incurred along data acquisition and preprocessing.

To address these limitations, *discriminative biclustering* algorithms have been proposed with specific criteria of interest to affect the structure, coherence and quality of the desirable patterns [26]. This is commonly guaranteed through the use of a merit function (such as the variance of the values in the bicluster) to guide the search. Following the taxonomy proposed by Madeira and Oliveira [26], biclustering algorithms can be characterized by their search paradigm, which determines how merit functions are applied. Greedy iterative searches rely on the selection, addition and removal of rows and columns until the merit function is maximized locally [27]. Exhaustive searches use merit functions to guide the space exploration [54]. Approaches combining clusters from both dimensions place merit functions for the clustering and joining stages [55]. Divide-and-conquer searches exploit the matrix recursively using a global merit function [56]. Stochastic approaches derive biclusters from multivariate distributions [19] and learn their parameters by maximizing a likelihood (merit) function. Biclustering can be easily extended for associative classification by defining class-conditional searches and adequately scoring the discriminative power of biclusters. Discriminative biclustering methods have been recently proposed for biomedical data analysis with different scores, including FDCluster [57], DRCluster [8], among others [17, 38]. Di-RAPOCC [13] considers a bicluster to be discriminative if it has high confidence and low inter-class overlapping. The major problems with the traditional discriminative biclustering approaches are two-fold:

L1. restrictions on the allowed number, positioning and quality of biclusters [15], causing associative classifiers to miss relevant subspaces;

L2. associative classifiers are unable to find non-constant patterns.

Discovery of non-constant patterns. Although many biclustering algorithms emerged in the last decade to find non-constant patterns (as motivated in Table 2), most algorithms still suffer from key limitations (see Table 3).

Coherence	*State-of-the-art algorithms*	*Limitations*
Additive and Multiplicative	Major attempts rely on merit functions based on variance, either more suitable to model additive factors (including residue-based approaches [58, 59]), or multiplicative factors (such as Fabia [19]). Some approaches unify these seemingly incompatible factors using linear geometry in hyper-spaces [60], evolutionary computing [61], and swarm intelligence [62].	1) Higher propensity to discover noisy constant biclusters instead of biclusters with strict additive and multiplicative coherence [15]; 2) Restrictions on the structure and quality of solutions.
Order Preserving	Greedy approaches iteratively discover-and-mask biclusters, including the pioneer OPSM [27] and its extensions to handle uncertainity [63]. Contrasting, few exhaustive approaches, such as uClustering [24], identify the largest subspaces that respect ranking order constraints, overcoming the quality and flexibility issues of the greedy peers.	1) Greedy solutions with restrictions on the structure (no overlaps) and no optimality guarantees; 2) Exhaustive approaches have efficiency bottlenecks and are highly susceptible to noise (perfect orderings only).

Table 3: Contributions and limitations of algorithms to discover non-constant biclusters.

A recent class of biclustering algorithms – pattern-based biclustering – can be used to tackle these limitations, opening a new door to study the impact of using non-constant coherencies for associative classification. State-of-the-art pattern-based biclustering algorithms rely on largely researched principles from pattern mining to guarantee an exhaustive yet efficient exploration of the search space with parameterizable coherency and quality [64, 28, 15]. In this context, frequent itemset mining, association rule mining, sequential pattern mining and graph mining can be applied to find constant, noisy, order-preserving and dense biclusters, respectively [64]. Principles can be placed to guarantee the scalability of these pattern mining searches under guarantees of optimality [28]. BicPAM [15], BicNET [21], BicSPAM [28] and Bic2PAM [65] extend the original pattern-based biclustering approaches [66, 67] (prepared to model constant coherencies) to further find additive, multiplicative and order-preserving biclusters robust to noise. Table 4 synthesizes their properties, while Fig.3 provides an illustrative result of their application. Recently, BicPAMS [2] was proposed to integrate state-of-the-art pattern-based biclustering algorithms, exploring their synergies.

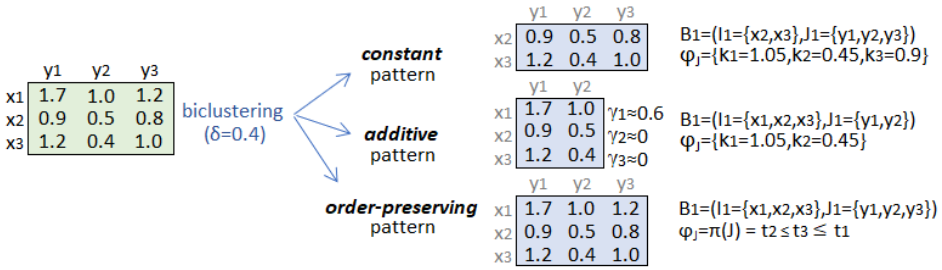

Figure 3: Integrative biclustering solution: search for biclusters with varying coherence.

Coherence and *method*	*Behavior*	*Tackled Limitations*
Constant BicPAM [15] BicNET [21]	BicPAM and BicNET integrate dispersed contributions from pattern-based biclustering approaches to offer the unprecedented possibility to customize the desirable properties of biclustering solutions, including their coherence strength, structure and tolerance to noise and missings. These algorithms may further incorporate domain knowledge [65].	Robustness to noise; Exhaustive (yet efficient) searches; Flexible structures; Extensible to multi-class settings.
Additive and Multiplicative BicPAM [15]	BicPAM makes use of variable-specific differences (additive case) and least common divisors (multiplicative case) between observations in order to perform iterative corrections for remove shifting and scaling factors. Pruning strategies are considered to avoid redundant calculus and reduce the computational complexity of these searches.	Solutions with flexible structure; Parameterizable quality; First exhaustive search to find shifting and scaling factors across observations.
Order Preserving BicSPAM [28] BicNET [21]	Patern-based searches can be parameterized with sequential pattern mining for an exhaustive discovery of biclusters with noise-tolerant orders. For this aim, the indexes of the features are reordered according to their values per observation and the biclusters are mapped from the frequent subsequences. BicSPAM and BicNET further allow a parameterizable variation of the degree of co-occurrences (elements with similar values) versus precedences to tune the desirable properties of the order-preserving coherence.	Tackled noise-intolerance and efficiency bottlenecks of exhaustive approaches; Flexible structures with guarantees of optimality, addressing problems of greedy approaches.

Table 4: Recent breakthroughs on biclustering (tackling limitations from Table 3).

4. Solution: Upgrading Associative Classification

With the aim of assessing the impact of enriching associative classifiers with non-constant patterns, this section introduces a new associative classifier, FleBiC (Flexible Biclustering-based Classifier). FleBiC explores the recent breakthroughs on the discovery of non-constant patterns (Table 4), and further addresses general limitations of existing associative classifiers [12, 10], including i) scarcity of matches, ii) intolerance to noise, iii) patterns without guarantees of statistical significance, iv) uneven space exploration, and v) generalization difficulties. Figure 4 summarizes the proposed contributions. Accordingly, FleBiC is driven by the following requirements:

R1. effective discovery of discriminative patterns (*discovery*);

R2. effective scoring and composition of discriminative patterns (*training*);

R3. effective matching of observations against scored patterns (*testing*).

In accordance with these requirements, FleBiC: 1) discovers flexible structures of discriminative biclusters with diverse coherence and quality (*Section 4.1*); 2) relies on state-of-the-art training functions with revised scoring schema to weight coherence type and strength (*Section 4.2*); and 3) defines testing functions tolerant to noise and able to match new observations against non-constant biclusters (*Section 4.3*). The pseudocode of FleBiC is provided in *Section 4.4*. Figure 5 summarizes the behavior of FleBiC.

4.1. Discovery of discriminative patterns

The state-of-the-art pattern-based biclustering algorithms surveyed in Table 2 were recently integrated in BicPAMS (Biclustering based on PAttern

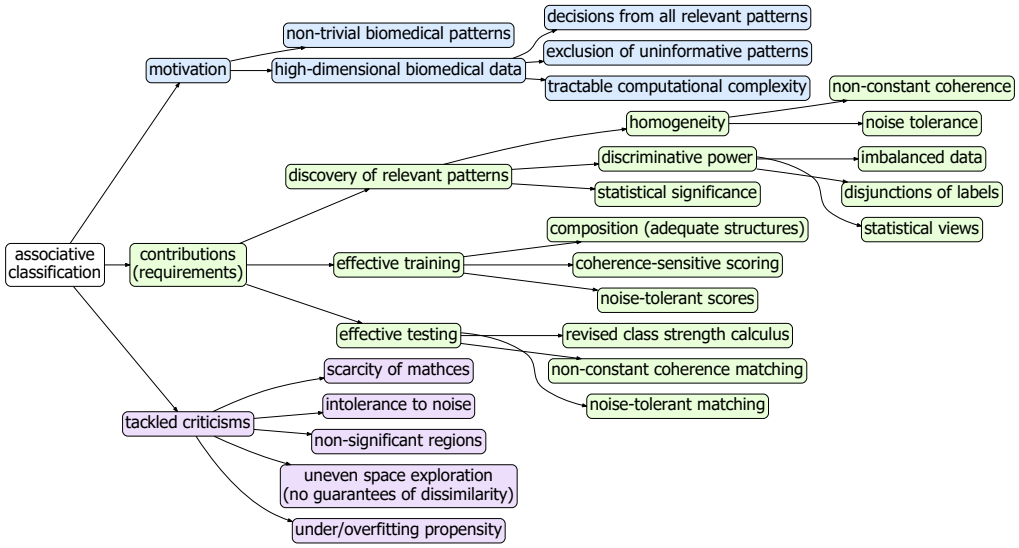

Figure 4: Structured view on the target associative classifiers: relevance, contributions and tackled criticisms. The FleBiC classifier incorporates all these principles.

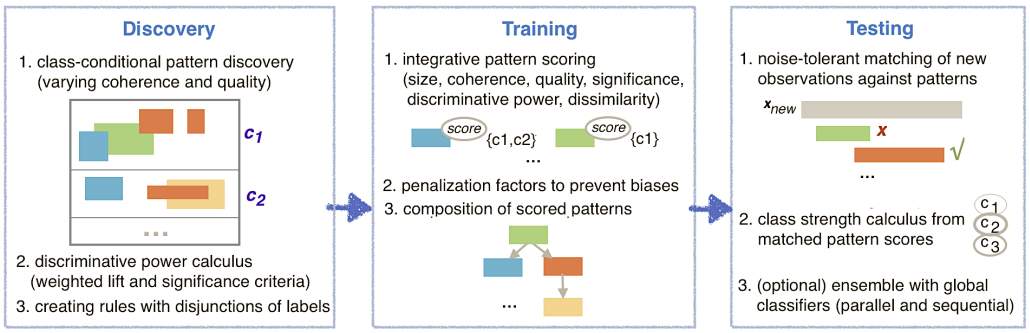

Figure 5: Summary of FleBiC behavior: *discovery* (biclustering with parameterizable homogeneity and generation of rules with disjunctive labels), *training* (scoring, penalization and composition of patterns) and *testing* (pattern matching and class strength calculus).

Mining Software) [2] and provide two major properties of interest: 1) the unprecedented possibility to discover biclusters with parameterizable coherence and quality, and 2) efficient searches with optimality guarantees. The exhaustive nature of BicPAMS, without restrictions on the number, size, positioning and homogeneity of biclusters, provide the unique possibility to:

- consider all potentially relevant patterns to support decisions, thus minimizing the criticisms of existing associative classifiers, namely underfitting propensity, and scarcity of matches;
- find noise-tolerant patterns with guarantees of statistical significance;
- assess the relevance of non-constant patterns on classification tasks.

Below we enhance BicPAMS – originally prepared for unsupervised learning tasks – to find patterns with guarantees of discriminative power.

Discovery of non-constant and noise-tolerant patterns. FleBiC applies BicPAMS [2] for each class-conditional data partition, returning $|\mathcal{C}|$ sets of biclusters. BicPAMS is applied with:

- constant, additive, multiplicative, and order-preserving assumptions;
- varying coherence strength on numeric variables, $\delta \in \{\frac{1}{3}\bar{\mathcal{Y}}_i, \frac{1}{4}\bar{\mathcal{Y}}_i, \frac{1}{5}\bar{\mathcal{Y}}_i, \frac{1}{7}\bar{\mathcal{Y}}_i, \frac{1}{10}\bar{\mathcal{Y}}_i\}$ where $\bar{\mathcal{Y}}_i$ is the amplitude of the domain of variable y_i;
- parametrically placed expectations on the quality of biclusters [2] (up to $\theta \in \{0, 10\%, 25\%\}$ noisy elements);
- statistical significance guarantees at α=1E-3 [68];
- remaining default parameters of BicPAMS (discussed in [2]) preserved.

BicPAMS is able to tackle the inherent computational complexity of performing numerous searches to satisfy the aforementioned homogeneity criteria (multiple coherence assumptions, coherence strength and quality thresholds) [2] by building upon intermediate biclustering solutions. Figure 6 illustrates the major steps undertaken for the (class-conditional) discovery of biclusters.

Support of non-constant and noisy patterns. The traditional notion of pattern support corresponds to the number of observations that perfectly respect a given pattern,

$$\sum_{\mathbf{x}_i \in \mathbf{X}} (\mathbf{x}_i \vdash \varphi_{\mathbf{J}}). \tag{1}$$

Def. 4.1 *Given a real-valued dataset with a bicluster* $\mathbf{B}=(\mathbf{I},\mathbf{J})$ *with coherence strength* δ *and pattern* $\varphi_{\mathbf{J}}$*, an observation* $\mathbf{x}_i$ perfectly respects $\varphi_{\mathbf{B}}$*, denoted* $\mathbf{x}_i \vdash \varphi_{\mathbf{B}}$*, if* $\forall_{a_{ij}\in(\mathbf{x}_i,\mathbf{J})}\ \eta_{ij}\in[-\delta/2, \delta/2]$ *(in accordance with* Def.2.4*). Given a symbolic dataset and bicluster* $\mathbf{B}$*, an observation* $\mathbf{x}_i$ perfectly respects $\varphi_{\mathbf{B}}$ *if* $\forall_{a_{ij}\in(\mathbf{x}_i,\mathbf{J})}\ \eta_{ij}=0$.

Given a dataset, and a bicluster $\mathbf{B}=(\mathbf{I},\mathbf{J})$ *with coherence across rows and pattern* $\varphi_{\mathbf{J}}$*, the* ***exact support*** *of the pattern* $\varphi_{\mathbf{J}}$*,* $sup_{\varphi_{\mathbf{J}}}$*, is given by (1), the number of observations perfectly respecting* $\varphi_{\boldsymbol{J}}$.

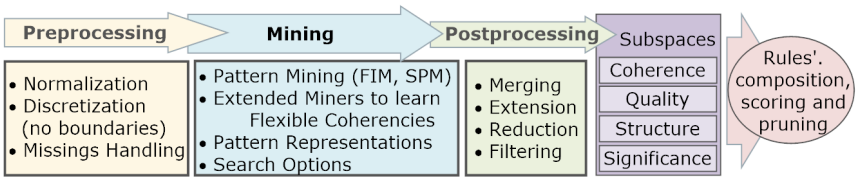

Figure 6: BicPAMS: major steps for the class-conditional discovery of non-constant and noise-tolerant biclusters.

This concept is not only valid for constant patterns but also for non-constant patterns (whether they follow additive, multiplicative, or order-preserving assumption) in real-valued datasets after removing the adjustment factors γ_i (in accordance with Defs. 2.4-2.5).

Illustrating, given an integer data space with pattern $\varphi_{\mathbf{J}}$=$\{k_2$=3,k_3=4, k_5=0$\}$, the $\mathbf{x}_2$ observation with a_{22}=4, a_{23}=5 and a_{25}=1 perfectly respects this pattern under an additive assumption (γ_2=1), and $\mathbf{x}_6$ observation with a_{62}=2, a_{63}=5 and a_{65}=1 perfectly respects this pattern under an order-preserving assumption, $\pi(\mathbf{J})$=$y_5 \leq y_2 \leq y_3$.

Despite its relevance, the traditional notion of support cannot account for deviations from pattern expectations in the presence of noise. For instance, consider a constant pattern $\varphi_{\mathbf{J}}$=$\{k_3$=0.4,k_5=0.9,k_7=0.1$\}$ and an observation $\mathbf{x}_2$ with a_{23}=0.4, a_{25}=0.8 and a_{23}=0.1. Given δ<0.1, the observation $\mathbf{x}_2$ does not support the pattern $\varphi_{\mathbf{J}}$ since $\eta_{25} \notin [-\delta/2, \delta/2]$ (η_{25}=0.1). As illustrated by this example, the exact support cannot account for the presence of localized forms of noise. In this context, in order to correctly weight the effect of noise, we propose a revised notion of support (Def.4.2) by assessing if a given observation is satisfied a noise threshold (percentage of values not satisfying a given coherence strength below the predefined threshold).

Def. 4.2 *Given a bicluster* $\mathbf{B}$=$(\mathbf{I}, \mathbf{J})$ *with coherence strength* δ*, the* matching factor κ_i *between observation* $\mathbf{x}_i$ *and the pattern* $\varphi_{\mathbf{J}}$ *is*

$$\kappa_i = \frac{1}{|\mathbf{J}|} \sum_{a_{ij} \in (\mathbf{x}_i, \mathbf{J})} (\eta_{ij} \in [-\delta/2, \delta/2] \wedge y_i \in \mathbb{R}) \vee (\eta_{ij} = 0), \tag{2}$$

corresponding to the fraction of variables in $\mathbf{J}$ *respecting value expectations.*

Given a dataset and a bicluster $\mathbf{B}$=$(\mathbf{I}, \mathbf{J})$ *with coherence across rows and pattern* $\varphi_{\mathbf{J}}$= $\{k_j \mid \mathbf{y}_j \in \mathbf{J}\}$*, the* ***weighted support*** *of the pattern* $\varphi_{\mathbf{J}}$ *is*

$$\text{wsup}_{\varphi_{\mathbf{J}}} = \sum_{\mathbf{x}_i \in \mathbf{X} \wedge \kappa_i < \epsilon} (\kappa_i)^a, \tag{3}$$

where κ_i *is the matching factor,* ϵ *is the minimum match threshold, and* a *is the noise penalization factor.*

Illustrating, given a constant pattern $\varphi_{\mathbf{J}}$=$\{k_3$=0.4,k_5=0.9,k_7=0.1$\}$ and δ<0.1, the observation $\mathbf{x}_2$ with a_{23}=0.4, a_{25}=0.8 and a_{23}=0.1 has a κ_2=$\frac{2}{3}$ matching factor against $\varphi_{\mathbf{J}}$, contributing to the support of $\varphi_{\mathbf{J}}$ with κ_2^2=$\frac{2}{3}^2$=0.(4).

By changing the noise-controlling factors a and ϵ in (3), the weighted support of a pattern can be adjusted to the level of noise according to sublinear, linear or squared (default) penalizations (Def.4.2). From empirical evidence (details in Section 4.4), ϵ=0.6 and a=2 are the suggested default values.

Similarly to the classic support (Def.4.1), the weighted support is equally applicable to non-constant patterns by detecting and removing the adjustment factors γ_i (in accordance with Defs.2.4-2.5).

Discriminative patterns. Discovering class-conditional biclusters with high support does not necessarily imply that they are discriminative if their support is high for all classes. In order to assess and guarantee the discriminative power of the (class-conditional) biclusters, FleBiC uses the introduced weighted notion of support to revise scores from information gain theory and performs additional statistical tests.

Def. 4.3 *Given a set of labels C in $\mathcal{C}$, the support of C, sup_C, is the number of observations with a label in C, $sup_C = |\{(x_i, c_i) \mid i = 1..n \wedge c_i \in C\}|$.*

Given a set of observations $\mathbf{X}$ and an association rule $R : \varphi_{\mathbf{J}} \Rightarrow C$:

*The **weighted support** of a rule is the weighted pattern support (according to Def.4.2) for observations with a label in C,* $\text{wsup}_R = \{\text{wsup}_{\varphi_{\mathbf{J}}} \mid \mathbf{x}_i \in C\}$.

*The **weighted confidence** of a rule is thus*

$$\text{wconf}_R = \frac{\text{wsup}_R}{\text{wsup}_{\varphi_{\mathbf{J}}}}, \tag{4}$$

*and its **weighted lift** is*

$$\text{wlift}_R = \frac{\text{wsup}_R}{\text{wsup}_{\varphi_{\mathbf{J}}}\text{sup}_C}. \tag{5}$$

FleBiC uses three discriminative indicators by default:

- **weighted lift** (5): the weighted lift extends the classic lift (originally proposed in the context of transactional data analysis) towards noise-tolerant patterns possibly following non-constant coherencies. Lift reveals the discriminative power of a pattern in accordance with the representativity of the labels in the consequent. Thus, it is more appropriate than confidence for imbalanced data.
- **statistical significance**: recent work on the statistical significance of biclustering solutions [68] provides statistical tests to assess biclusters with constant, additive, multiplicative, symmetric and order-preserving assumptions. These statistical tests deliver a (corrected) p-value defining the probability of a (possibly non-constant) pattern support to deviate from expectations [68]. In this context, and given an association rule $R : \varphi_{\mathbf{J}} \Rightarrow C$, FleBiC assesses whether the bicluster with $\varphi_{\mathbf{J}}$ is statistical significant for the observations with label in $C \subset \mathcal{C}$ (p-value lower than 0.01) and not significant for the observations with label in $\mathcal{C} \setminus C$ (p-value higher than 0.05).
- χ^2 **test**: complementary view of a pattern's discriminative power placed by peer state-of-the-art associative classifiers [29].

Disjunctions of labels. Since specific patterns may not be able to discriminate a single label but instead discriminate a set of labels, we allow for disjunctions of labels in the consequents of the rules. Illustrating, given three classes $\{c_1, c_2, c_3\}$, a pattern that is only statistically significant for observations with c_1 or c_2 is able to discriminate $\{c_1, c_2\}$ from c_3.

As we allow for disjunctions of labels in the consequent, weighted lift is preferable over weighted confidence to deal with the imbalance incurred from grouping labels in the consequent.

FleBiC efficiently generates rules with disjunctive sets in the consequent by analyzing the statistical significance and discriminative power of a pattern for each class. If a pattern $\varphi_{\mathbf{J}}$ is statistically significant on multiple classes $c \in C$ (where $C \subset \mathcal{C}$) – $\forall_{c \in C} P(\varphi_{\mathbf{J}}|c)$<1E-3 – yet not able to discriminate each class individually, a new rule $\varphi_{\mathbf{J}} \Rightarrow C$ is generated if $P(\varphi_{\mathbf{J}}|C)$<1E-3 and $P(\varphi_{\mathbf{J}} \mid \mathcal{C} \backslash C)$>5E-2.

Since statistically testing biclusters can be performed in linear time [68], the computational complexity of merging association rules to allow for disjunctions of labels is polynomial on the number of patterns and classes.

4.2. Training

Given a set of patterns, the subsequent question is how to adequately score and organize them given their diverse size, coherence and quality.

Integrative scoring. Scoring is key in associative classification since it defines the ability of a given pattern to discriminate a subset of classes. Effective scoring is also relevant to tackle problems associated with the imbalanced number of patterns per class or the overemphasis on small patterns.

To guarantee a scoring that accounts for all these variables, we propose an integrative score combining the rule's discriminative power (using the proposed weighted lift, discriminative significance and χ^2 test) with additional four indicators: 1) pattern length, 2) pattern weighted support, 3) quality (deviation from the pattern expectations, η_{ij}), and 4) statistical significance [68]. This overcomes the typical problem of associative classifiers that prioritize small (often non-significant) biclusters as a result of an overemphasized focus on the confidence of the rules.

Def. 4.4 *Given a dataset with an association rule* $R : \varphi_{\mathbf{J}} \Rightarrow C$, *let* $T_{sig(R)}$ *be an assessment of the statistical significance of* R: *1 if* $P(\varphi_{\mathbf{J}}|C)<0.01 \wedge P(\varphi_{\mathbf{J}}|C)>0.05$ *and 0.1 otherwise. Let* Q_B *be the quality of a bicluster* $\mathbf{B}=(\mathbf{I}, \mathbf{J})$ *given by the fraction of non-noisy elements,*

$$Q_B = \frac{1}{|\mathbf{I}|} \sum_{\mathbf{x}_i \in \mathbf{I}} \kappa_i, \tag{6}$$

where κ_i *is defined according to Def.4.2.*

In this context, the ***integrative score*** *of a rule* $R : \varphi_{\mathbf{J}} \Rightarrow C$ *is*

$$\omega_R = T_{disc(R)} \times \omega'_R, \quad with \tag{7}$$

$$\omega'_R = \alpha_1\big(0.7\frac{\text{wsup}_R}{\text{wsup}_{\varphi_B}}\frac{sup_C}{sup_{\mathcal{C}}} + 0.3\chi^2_{\varphi_B}\big) + \alpha_2\big(0.5\frac{\text{wsup}_R}{n}\frac{sup_C}{sup_{\mathcal{C}}} + 0.5\frac{|\varphi_B|}{m}\big) + \alpha_3 Q_B, \tag{8}$$

where, from empirical evidence, α_1*=0.6,* α_2*=0.3 and* α_3*=0.1 by default (details provided in Section 4.4).*

To illustrate the proposed integrative score, consider the data in Table 5, the shaded bicluster $\mathbf{B}_1$ with pattern $\varphi_{\mathbf{J}_1}$=$\{k_4$=2,k_7=4,k_9=3, k_{20}=1$\}$, and rule $R_1 : \varphi_{\mathbf{J}_1} \Rightarrow \{c_2, c_3\}$. In this context, n=$|\mathbf{X}|$=9, m=$|\mathbf{Y}|$=20, $\text{sup}_{\mathcal{C}}$=n=9, sup_C=$\text{sup}_{\{c2,c3\}}$=5, $\text{sup}_{\varphi_{\mathbf{J}_1}}$=$|\{x_5, x_7\}|$=2, $\text{wsup}_{\varphi_{\mathbf{J}_1}}$=2+3×$0.75^2$=3.69, sup_{R_1}=2, wsup_{R_1}=3.13, $\chi^2_{\varphi_{\mathbf{J}_1}}$=$\frac{2.72}{9}$=0.3, Q_{R_1}=$\frac{4\times4-2}{4\times4}$=0.875, and $T_{sig(R_1)}$=1 since $P(\varphi_{\mathbf{J}_1}|C)$=$\binom{m}{|\varphi_{\mathbf{J}_1}|}\Sigma_{x=sup_{\varphi_{\mathbf{J}_1}}}^{n}\binom{n}{x}p_{\varphi_{\mathbf{J}}}{}^{x}(1\text{-}p_{\varphi_{\mathbf{J}}})^{n-x}$=1.6E-7 and $P(\varphi_{\mathbf{J}_1}|\mathcal{C}/C)\approx$1 [68]. As such, ω_{R_1}=$T_{sig(R_1)}$×ω'_{R_1}=1×(0.6×(0.7×0.47+0.3×0.3)+0.3×(0.5×0.193+0.5×0.2)+0.1×0.875)=0.39. Identically, the rule associated with bicluster $\mathbf{B}_2$, $R_2 : \varphi_{\mathbf{J}_2} \Rightarrow \{c_1\}$ where $\varphi_{\mathbf{J}_2}$=$\{k_4$=3,k_7=1,k_8=5,k_{20}=4$\}$ has ω_{R_2}=0.05.

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

$\mathbf{B}_1 = (\mathbf{I}_1 = \{x_3, x_5, x_6, x_7, x_9\}, \mathbf{J}_1 = \{y_4, y_7, y_9, y_{20}\})$

$\mathbf{B}_2 = (\mathbf{I}_2 = \{x_1, x_4\}, \mathbf{J}_2 = \{y_4, y_7, y_8, y_{20}\})$

Table 5: Illustrative dataset in $\mathbb{N}_0^+$ with two highlighted biclusters.

Pattern dissimilarity. FleBiC penalizes patterns that are similar to other available pattern that discriminates the same classes with a higher integrative score. Dissimilarity is defined by the number of non-shared variables, being the penalization given by its square root (from empirical evidence).

Illustrating, given two patterns with $|\varphi_{\mathbf{J}_1}|$=$|\varphi_{\mathbf{J}_2}|$=4, $\omega_{\mathbf{J}_1 \Rightarrow C} < \omega_{\mathbf{J}_2 \Rightarrow C}$ and three shared variables, then the score $R_1 : \varphi_{\mathbf{J}_1} \Rightarrow C$ is $\sqrt{(1 - \frac{\mathbf{J}_1 \cap \mathbf{J}_2}{\mathbf{J}_1})}\omega_{R_1}$=$\sqrt{\frac{1}{4}}\omega_{R_1}$.

Scoring non-constant patterns. The introduced integrative score does not address the fact that different coherence assumptions may show different degrees of flexibility. As illustrated in Figure 3, order-preserving biclusters have higher flexibility degree as they are able to capture additive and

multiplicative patterns, which in turn are able to capture constant patterns. Understandably, coherencies with higher flexibility (typically associated with larger biclusters) will have higher scores and can jeopardize the learning since biclusters given by more restrictive coherencies become neglected. For this reason, it is important to weight the score of patterns in accordance with their coherency assumption.

To this end, we introduce a new penalization weight, $\omega \times \nu$, for non-constant coherencies based on their degree of flexibility. From empirical evidence, the following penalizations are provided by FleBiC as default: order-preserving (ν=0.7 with symmetries and ν=0.75 otherwise); additive (ν=0.8 with symmetries and ν=0.85 otherwise); multiplicative (ν=0.9); and constant with symmetries (ν=0.95).

Composition of patterns. Given a set of scored patterns, the possibility to traverse them efficiently is relevant for an efficient testing of new observations. The simplest option is to simply rely on an ordered set of tuples (pattern $\varphi_{\mathbf{K}}$, labels C, score $\omega_{\varphi_{\mathbf{J}} \Rightarrow C} \times \nu$). To guarantee a better navigation throughout the patterns, FleBiC uses the tree structure proposed by CMAR [29], where rules are organized according to their score, consequent and pattern length.

4.3. Testing

In the testing stage, the learned associative model is used to label a new observation $\mathbf{x}_{new}$ by: 1) identifying matching patterns per class; and 2) computing the class strength, $\forall_{c \in \mathcal{C}} P(c \mid \mathbf{x}_{new})$, based on the extent and score of the matched patterns.

Noise-tolerant matching of patterns. Given an association rule, $R : \varphi_{\mathbf{J}} \Rightarrow C$ where $C \subset \mathcal{C}$, exact matching occurs if the values of the testing observation $\mathbf{x}_{new}$ respect a pattern, $\mathbf{x}_{new} \vdash \varphi_{\mathbf{J}}$ (Def.4.2).

Yet, even in the presence of a large number of patterns, the probability of matches to occur can be considerably low. Thus, the introduction of relaxations is critical to consider matches when a testing observation: 1) respects the majority (but not all) of the expected values of a pattern, or 2) the overall noise is below a given threshold.

Def. 4.5 *Given a rule $R : \varphi_{\mathbf{J}} \Rightarrow C$ with score $\omega_R \times \nu$ and a matching threshold θ, an observation $\mathbf{x}_{new}$* ***matches*** *$\varphi_{\mathbf{B}}$ if it respects $\varphi_{\mathbf{B}}$ ($\kappa_{new} > \theta$, where κ_{new} is given by Def.4.2) with* ***matching score*** *$\omega_R \times \nu \times {\kappa_{new}}^a$, where θ=0.6 and a=2 from empirical evidence (sensitivity analysis detailed in Section 4.4).*

Given the rules R_1 and R_2 discussed in Table 5, an observation $\mathbf{x}_{new}$ with $a_{new,4}$=3, $a_{new,7}$=4, $a_{new,8}$=1, $a_{new,9}$=3 and $a_{new,20}$=1 matches $\varphi_{\mathbf{J}_1}$ with score $\omega_{R_1} \times \nu \times \kappa^2_{new|B_1}$=0.39×0.75^2 and does not match $\varphi_{\mathbf{J}_2}$, $\kappa_{new|\varphi_{\mathbf{J}_2}}$=0.25<0.6.

Matching non-constant patterns. To determine if a testing observation respects a non-constant pattern, we need to verify if the observed values can be described by an adjustment factor. Given a non-constant pattern $\varphi_{\mathbf{J}}$, an observation $\mathbf{x}_{new}$ *matches* $\varphi_{\mathbf{J}}$ if it respects $\varphi_{\mathbf{J}}$ assuming $\gamma \neq 0$ or the presence of ordering constraints. Illustrating, consider a bicluster with pattern $\varphi_{\mathbf{J}}$=$\{$1.2, 3.3, 2.0$\}$ on variables $\{y_{89}, y_{459}, y_{892}\}$. If the bicluster is additive and a testing observation has values $\{$3.1, 5.3, 4.1$\}$ on the same variables, the values are coherent under a shifting factor γ=2. If the same bicluster is order-preserving, $\mathbf{x}_{new}$ is also coherently described $\varphi_{\mathbf{J}}$=$\pi(\mathbf{J})$=$y_{89} \leq y_{892} \leq y_{459}$.

Class strength. In FleBiC, the strongest class, $c \in \Sigma$, is outputted as the estimated class, if we want a deterministic output. Otherwise, the strength of each class is normalized and returned.

Def. 4.6 *Given a new observation* $\mathbf{x}_{new}$ *and matched patterns* Φ*, the strength of a class* $c \in \mathcal{C}$ *is by default given by its* ***weighted integrative score****,*

$$WIS_c = \sum_{(\varphi_{\mathbf{J}} \Rightarrow C) \in \Phi \wedge c \in C} \frac{sup_c}{sup_C} \nu \times \omega_{\varphi_{\mathbf{J}} \Rightarrow C}. \tag{9}$$

Given the already covered properties of ω, the class strength calculus given by (9) is simplistic and able to accommodate rules with disjunctions in the consequent. Yet, it is empirically shown to be more effective than state-of-the-art alternatives [29, 42], such as the weighted-χ^2 (even when considering its extension to deal with disjunctions of labels on rules' consequent[1]).

Combining local and global views. Different testing observations may be classified with different degrees of confidence due to the extent of matches and the consistency of labels from the matched rules. FleBiC makes available strategies to tackle the following situations: *1)* few matched patterns or matched patterns with low scores, *2)* no label with significantly higher probability (weak consistency of rules' consequent), and *3)* observations not only characterized by local patterns but also by global regularities.

First, FleBiC can rely on the (probabilistic) output of other well-known classifiers able to focus on non-local data distributions, here referred as global classifiers. For instance, the output of support vector machines, Bayesian classifiers and multivariate discriminants can be considered due to their contrasting behavior against associative classifiers [69]. In this context, the (normalized) probability per label, $\mathbf{p}_L$=$\{P(c_1|\mathbf{x}_{new}), .., P(c_{|\Sigma|}|\mathbf{x}_{new})\}$, is weighted

[1] *weighted-*$\chi^2(c)$=$\Sigma_{\varphi_B \Rightarrow C \in P \wedge c \in C} \frac{sup_c}{sup_C} (\chi^2_{\varphi_B})^2$/MCS, where MCS=$(min(sup_{\varphi_B}, sup_C) - sup_{\varphi_B} sup_C/N)^2 N \times e$, N is the number of matches and e=$1/(sup_{\varphi_B} sup_C) + 1/(sup_{\varphi_B} N - sup_C) + 1/(N - sup_{\varphi_B} sup_C) + 1/(N - sup_{\varphi_B}(N - sup_C)$

with the output of d global classifiers ($\mathbf{p}_{Gi}$): $\mathbf{p}=\alpha\mathbf{p}_L+\frac{(1-\alpha)}{d}\Sigma_{i=1}^{d}\mathbf{p}_{Gi}$, where $\alpha\approx 0.4$ (from empirical evidence), by default.

Second, in the presence of matches but not a delineated preference towards a single label, the labels with significantly low probability to occur for a given testing observation can be excluded and not used as input to train the global classifiers. Contrasting, when there is less than $2\times|\Sigma|$ matches, the inverse strategy is considered: the C labels with lower probability from global classifiers are excluded from $\mathbf{p}_L$ calculus ($P(c \in C \mid \mathbf{x}_{new})=0$) in order to reduce the propensity towards unneccessary biases.

4.4. Algorithm

Algorithm 1 describes FleBiC. In the training stage, FleBiC relies on the class-conditional application of BicPAMS for the exhaustive discovery

Algorithm 1: FleBiC Core Steps

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

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

of coherent, dissimilar, statistically significant and discriminative patterns. Second, it generates rules associating these patterns with disjunctions of labels in the consequent, $R : \varphi_{\mathbf{J}} \Rightarrow C$ where $C \subset \mathcal{C}$, to minimize the problem of match scarcity in multi-class data contexts. Third, these rules are used to compose the classifier according to the proposed scoring schema and weights for balancing coherencies with distinct degree of flexibility. In the testing stage, the weighted integrated score is applied with relaxations to adequately match observations against non-constant patterns. In the presence of classification decisions with low-to-medium levels of confidence, the resulting score is integrated with the output of alternative classifiers.

Computational complexity. The computational complexity of FleBiC is bounded by the complexity of the biclustering task, which depends on the size and dimensionality of the class-conditional matrix, distribution of values, selected coherencies and merging procedure (details in [15, 28]). Scalability principles from pattern mining (data partitioning strategies and approximate searches) [28, 2] and domain constraints (based on user expectations and available background knowledge) [65] can be incorporated to promote the efficiency of the biclustering task. The training and testing steps are linear on the number of patterns and their average number of features.

FleBiC parameterization. Although parameterizable, FleBiC can be effectively applied with default parameterizations. In this scenario, BicPAMS is applied with multiple coherence assumptions and remaining default behavior (*Section 4.1*). The suggested parameters associated with the training and testing functions (Defs.4.4-4.6) were fixed according to a sensitivity analysis conducted for synthetic data (Table 6) by iteratively varying the values for each parameter (α_1, α_2, α_3, ν, θ and a) in order to maximize the harmonic mean of accuracy and sensitivity. Still, for the purpose of understanding and improving the performance of FleBiC for specific data domains, its behavior can be easily parameterized. To this end, FleBiC provides the distinct possibility to parameterize the coherence and quality of the target patterns, a lower bound on their statistical significance and discriminative power, as well as scoring-and-matching thresholds.

5. Results and Discussion

Results are organized as follows. *Section 5.1* compares the performance of FleBiC against state-of-the-art classifiers using synthetic and real biomedical data. *Section 5.2* deepens the analysis of FleBiC's performance, showing the relevance of learning from non-constant patterns. Results were gathered using a 10-fold cross-validation and differences statistically tested under a t-Student (9 degrees of freedom) and α=0.05 significance threshold. FleBiC is available

at http://web.ist.utl.pt/rmch/software/bclassifier. Experiments were run using an Intel Core i5 2.80GHz with 6GB of RAM.

Synthetic data. FleBiC makes available a new generator able to plant global regularities (class-conditional multivariate distributions) and local patterns. Some of the parameters that can be varied include:

- data size ($|\mathbf{X}| \times |\mathbf{Y}|$), number of classes, class imbalance, and a mixture of multivariate distributions to describe class-conditional data;
- number of biclusters, allowed coherencies, noise and associated overlapping/plaid effects [18];
- pattern's discriminative power, support and length (using both Uniform and Gaussian distributions).

The provided results are an average of 30 data instances per setting. Table 6 describes the parameters used to generate the synthetic datasets.

Sizes ($\|X\|\times\|Y\|$)	100×100	400×400	1000×1000	2000×5000
Number of biclusters	4	5	7	10
Bicluster length $\|J\|$	[5,10]	[10,20]	[15,30]	[25,50]
Absolute support $\|I\|$	[20,50]	[100,200]	[250,500]	[500,1000]

Settings for the 1000×1000 dataset (*default* in bold):

Number of classes $|\mathcal{C}| \in \{2,\mathbf{3},5\}$ with $\{0,\mathbf{0.2},0.5\}$ imbalance degree;
Coherence strength δ={5%,**10%**,25%,50%}×$\bar{A}$ and noisy elements {0%,**1%**,2%,5%,10%}
Coherence assumption $\mathcal{B}\sim$[{7 constant},{**3** additive, **2** multiplicative, **2** order-preserving}]
Discriminative power $\mu(\phi)$={**90%**,80%,70%,60%} and $\sigma(\phi)$={**2%**,4%}
Overlapping degree {0,**0.1**,0.2,0.5} with plaid effect f={**sum**,product} (according to [18])

Table 6: Properties of the generated synthetic data.

Real data. Classifiers were further assessed on 12 datasets: 4 gene expression datasets and 8 clinical databases. The selected gene expression datasets are: *lymph* [70], *leukemia* [71], *embryo* [72] and *colon* [73] (accessible via the webpage of BIGS research group[1]). These datasets were originally proposed for the classification of: *1)* distinct types of lymphoma (n=96 observations, m=4026 variables); *2)* leukemia (n=72, m=7129); *3)* embryonal tumour outcome (n=60, m=7219); and *4)* colon cancer (n=62, m=2000).

In addition to high-dimensional gene expression data, the 8 selected clinical datasets are: *hepatitis*[2], *postoperative*[3], *lung*[4], *mammography*[5], *statlog*[6],

[1] http://eps.upo.es/bigs/datasets.html
[2] https://archive.ics.uci.edu/ml/datasets/Hepatitis
[3] https://archive.ics.uci.edu/ml/datasets/Post-Operative+Patient
[4] https://archive.ics.uci.edu/ml/datasets/Lung+Cancer
[5] https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass
[6] https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29

hungarian[7], *cleveland*[10] and *breast*[8] (accessible in the UCI machine learning repository [74]). Table 7 synthesizes their major statistics, including number of classes, size, dimensionality, types of variables and purpose. In contrast with the synthetic and biological data, these datasets are not high-dimensional and thus offer the opportunity for a broader assessment of FleBiC.

	hepatitis	postoperative	lung	mammography	heart	hungarian	cleveland	breast
	$\|\mathcal{C}\|$=2 n=155 m=20	$\|\mathcal{C}\|$=3 n=90 m=9	$\|\mathcal{C}\|$=3 n=32 m=57	$\|\mathcal{C}\|$=2 n=961 m=5	$\|\mathcal{C}\|$=2 n=270 m=14	$\|\mathcal{C}\|$=2 n=294 m=14	$\|\mathcal{C}\|$=5 n=303 m=14	$\|\mathcal{C}\|$=6 n=106 m=10
Variables	Nominal	Nominal with missings	Mix (real, binary)	Nominal with missings	Mix (nominal, ordinal, real)	Mix (nominal, ordinal, real)	Mix (nominal, ordinal, real)	Real
Outcome	Die or live	Intensive care, mid-care or home	3 types of lung cancer	Benign vs. malignant tumor	Presence of heart disease	Coronary artery stenosis	Angiographic pathologies	Carcinoma, connective, mastopathy, glandular, fibro-adenoma, adipose

Table 7: Properties of the selected clinical datasets.

5.1. Comparison with state-of-the-art classifiers

Figures 7, 8 and 9 compare the performance of FleBiC against state-of-the-art classifiers over biological, clinical and synthetic data respectively. We compared associative classifiers based on pattern mining (using CMAR [29] after discretizing data using the suggested coherence strength thresholds) and biclustering (using FDCluster [57]), as well as support vector machines (SVM) and Bayesian networks (BayesNet) from WEKA [75]. For the sensitivity calculus, we consider the positive class to be associated with the observations having the pathology. Generally, the collected results show significant accuracy gains (after t-testing differences), positioning FleBiC as a promising approach for the analysis of high-dimensional data.

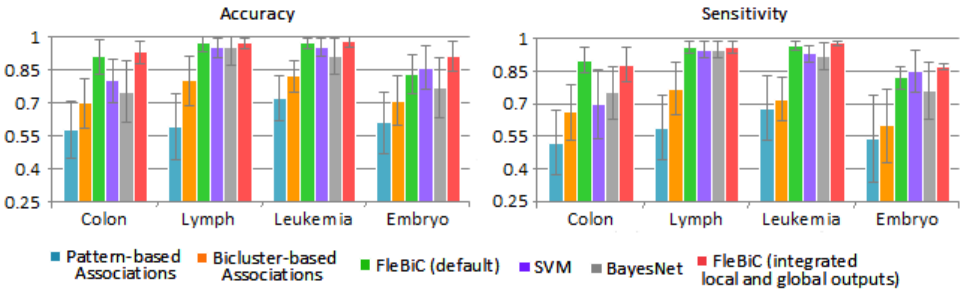

Figure 7: Comparison of FleBiC's performance against state-of-the-art associative and global classifiers over gene expression data.

[7] https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

[8] https://archive.ics.uci.edu/ml/datasets/Breast+Tissue

Results collected in Figures 7 and 8 highlight the superiority of FleBiC against peer pattern-based classifiers for the targeted biomedical datasets, confirming the importance of relying on comprehensive discovery of patterns with varying coherence and quality. BlueSecond, results further show that the accuracy and F-measure of FleBiC is generally competitive against state-of-the-art classifiers, and superior in some datasets, including *lung*, *colon*, *leukemia*, or *breast* (t-Test at α=0.05). This observation confirms the importance of focusing on local patterns. Third, sensitivity estimates further show the ability of FleBiC to handle imbalanced data. Finally, these analyzes also quantify the gains from integrating the output of FleBiC with well-known classifiers (as described in *Section 4.3*).

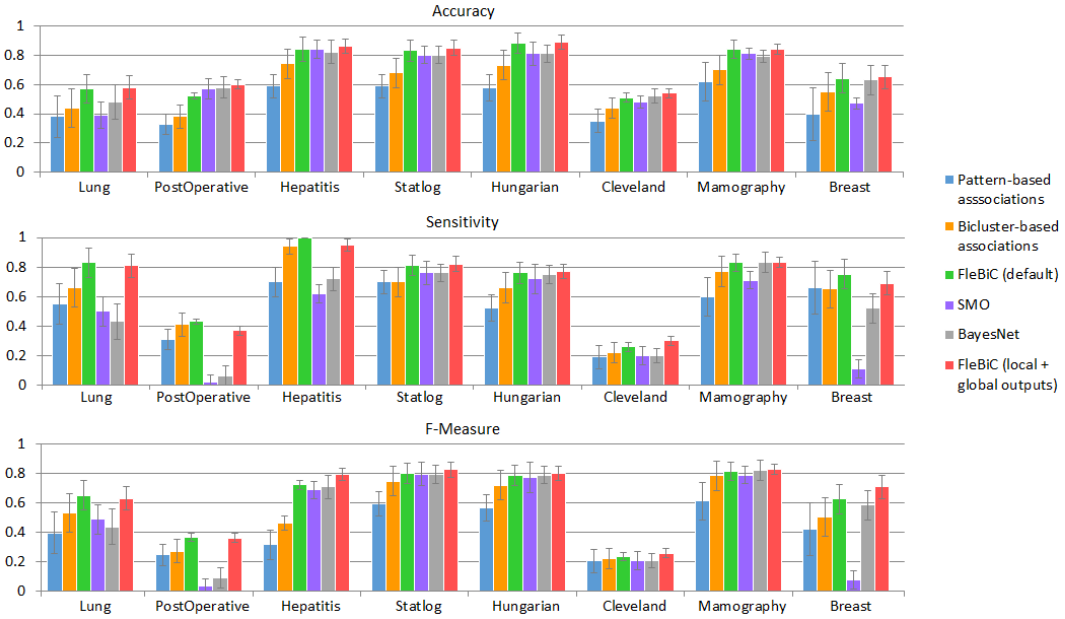

Figure 8: Comparison of FleBiC's performance over clinical data (Table 7) against associative classifiers and global classifiers. Sensitivity and F-measure computed for the minor class (e.g. glandular cancer in breast data) and home care in postoperative data.

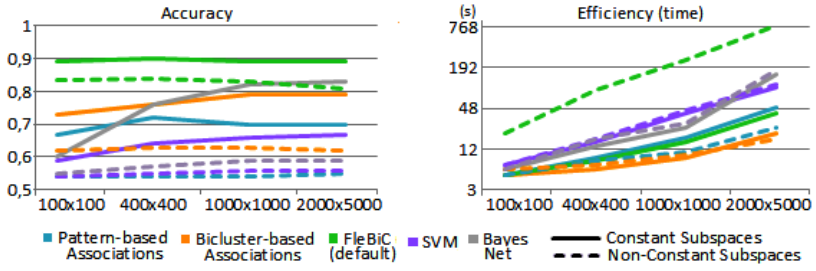

Figure 9: Accuracy and efficiency levels of FleBiC against state-of-the-art classifiers over synthetic data (Table 6).

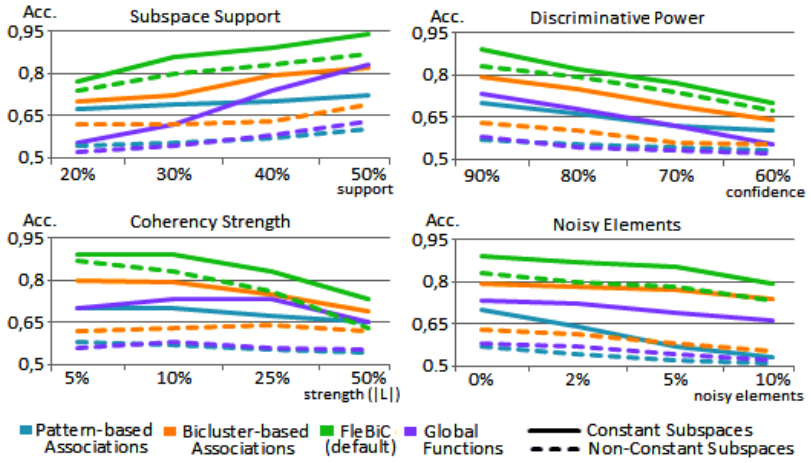

Figure 10: FleBiC's ability to learn from discriminative biclusters against peer classifiers in the presence of patterns with varying support, discriminative power, coherency type and strength, and noise.

Results from synthetic data (Figures 9 and 10) further confirm these observations. These analyses show that state-of-the-art classifiers are not well-prepared to learn from data with discriminative yet non-constant patterns, showing that there is still space to improve state-of-the-art classifiers.

The levels of efficiency in Figure 9 show that, although FleBiC's efficiency is penalized by the increased complexity associated with the discovery of patterns with diverse coherence, it is scalable (even in the presence of high-dimensional data). Figure 10 validates whether FleBiC is able to accurately classify synthetic data with non-constant patterns when varying their: coherence strength, number of supporting observations, amount of noise and discriminative power (Table 6). Results show FleBiC distinct superiority against associative classifiers (CMAR [29] and FDCluster [57]) and global classifiers (support vector machines, Bayesian networks and multivariate discriminants), motivating the importance of non-constant patterns tolerant to noise and adequate scoring criteria.

Figure 11 measures the impact of parameterizing FleBiC with alternative training and testing criteria, showing the possibility to adapt FleBiC behavior without hampering usability due to its effective default behavior.

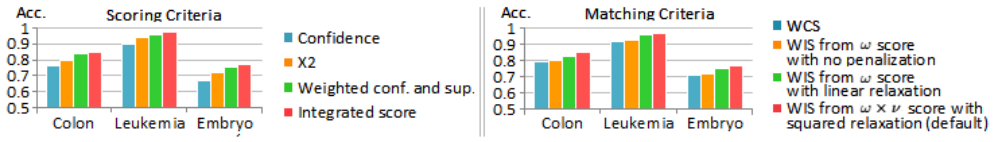

Figure 11: Impact of parameterizing scoring and matching functions when learning from gene expression data.

5.2. Relevance of non-constant patterns in biomedical data

Results presented along Figures 7-10 already evidence the relevance of learning from patterns with varying coherence and quality. This section deepens previous analyzes by quantifying the gains from considering non-constant assumptions, and studying their properties and biomedical relevance.

Regulatory Patterns. Table 8 motivates the importance of discovering non-constant patterns for biological data analysis. It measures the impact produced by each coherence assumption on the: percentage of confident decisions (over 10 matched patterns and a single class with distinctive higher probability), average pattern length and average rule weighted confidence (Def. 4.3). For this analysis, FleBiC was parameterized with δ=1/6 (after column-based data normalization), 70% quality, and decreasing support until at least 50 statistically significant rules per class are found. We observe that including non-constant patterns tolerant to noise is key to better discriminate classes (+20pp). The gains increase when moving from the isolate use of each coherence towards their integrated use (+10pp). This improved ability to discriminate classes seems to be correlated with the pattern length and higher correlation strength of rules.

Coherence	% of decisions with high confidence				Average weighted support				Average weighted confidence			
	Colon	*Lymph*	*Embryo*	*Leukemia*	*Colon*	*Lymph*	*Embryo*	*Leukemia*	*Colon*	*Lymph*	*Embryo*	*Leukemia*
Constant (baseline)	**0.51**	**0.68**	**0.59**	**0.50**	**23±3**	**13±2**	**40±6**	**24±2**	**0.82**	**0.94**	**0.81**	**0.92**
Noisy Constant	0.69	0.83	0.77	0.72	24±3	13±2	42±5	25±3	0.81	0.94	0.82	0.92
Symmetric	0.66	0.81	0.65	0.78	24±3	13±2	42±6	25±3	0.79	0.91	0.80	0.91
Additive	0.70	0.83	0.78	0.73	25±3	14±2	42±6	27±3	0.79	0.89	0.81	0.91
Multiplicative	0.69	0.81	0.77	0.68	24±3	14±3	40±5	25±2	0.79	0.88	0.80	0.90
Orde-Preserving	0.78	0.92	0.95	0.72	27±4	20±2	36±7	23±4	0.83	0.86	0.91	0.79
Plaid	0.70	0.81	0.79	0.72	22±3	13±2	39±5	25±2	0.80	0.90	0.81	0.90
Integrated (FleBiC)	**0.81**	**0.93**	**0.95**	**0.78**	**25±3**	**14±2**	**42±5**	**24±3**	**0.89**	**0.97**	**0.90**	**0.95**

Table 8: Impact of making decisions using non-constant biclusters from high-dimensional biological data. Results from top-100 rules, with 30% to 60% supporting class-conditional observations. Criteria: *1)* percentage of testing observations (from 10 cross-fold validation) with >10 matchings and clear preference towards a single class (>20pp difference of probabilistic outputs), *2)* the proposed weighted support, and *3)* weighted confidence.

The selected association rules identify discriminative regulatory patterns for the phenotype under classification. An extensive analysis of the biological relevance of non-constant patterns can be found in [15, 21, 18]. These studies show that the possibility to discover less-trivial putative modules is essential to accommodate meaningful variations between individuals (either associated with different responsiveness of genes, stage of disease progression or with the effects of drugs on gene expression).

Figure 12 assesses variations in FleBiC's performance over gene expression data when parameterizing FleBiC with different coherence assumptions.

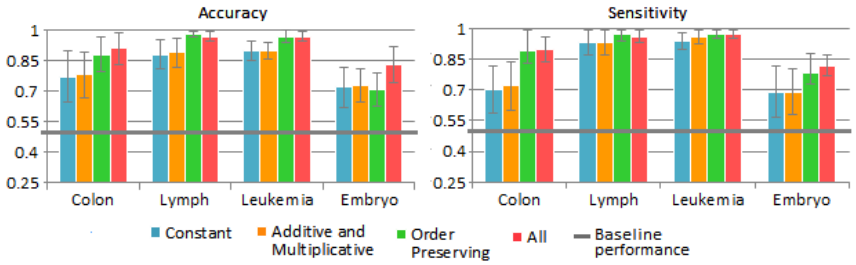

Figure 12: Accuracy and sensitivity gains of modeling non-constant biclusters from high-dimensional biological data.

Figure 12 shows significant improvements (p-value<0.05) in terms of accuracy and sensitivity for the integrated use of multiple coherencies, generally explained by an increased number of matches during the testing phase driven by the presence of a more diversified set of putative regulatory patterns.

The gathered results further highlight the significant role of order-preserving patterns in offering additional guidance to the behavior of associative classifiers. Improvements under the order-preserving assumption are statistically significant for the *colon*, *lymph* and *leukemia* datasets (p-value<0.05).

To complement these analyzes, Table 9 provides statistics associated with a few discriminative biclusters that share a single pattern $\varphi_{\mathbf{J}}$ but differ with regards to the quality (η_{ij}) and coherence assumption (γ_{ij}). The results of this analysis further show the importance of using noise-weighted criteria to better model the support of biclusters (and, consequently, to score rules), and the relevance of using non-constant patterns to increase the probability of matches and thus alleviate the common downsides of associative models.

data	$\varphi_{\mathbf{J}}$	$\lvert\varphi_{\mathbf{J}}\rvert$	*coherence*	ϵ-*noise* $\eta_{ij}<\epsilon$	*abs. support* $\{L,\bar{L}\}$	*weighted abs. sup.*	*integrative score*	*%test matches*
leukemia	P_1	9	constant	0.0	{9,0}	{10.3,0.9}	0.92	14%
leukemia	P_1	9	constant	0.2	{18,0}	{10.3,0.9}	0.92	14%
leukemia	P_1	9	additive	0.1	{19,0}	{13.4,2.1}	0.87	19%
colon	P_2	8	constant	0.0	{14,0}	{16.2,1.9}	0.89	11%
colon	P_2	8	plaid	0.1	{19,1}	{15.8,1.8}	0.90	13%
colon	P_2	8	order-preserving	0.1	{23,2}	{18.9,4.1}	0.86	18%

Table 9: Illustrative rules with fixed pattern $\varphi_{\mathbf{J}}$ (P_1={6,5.4,2.7,8.4,-3.6,6.3,5.1,8.1,7.5}, P_2={-3.3,-5.4,-5.7,-6,-2.7,-7.8,-8.1,-6.3}, $a_{ij}\in$[-10,10]) and varying coherence. Comparison of (weighted) absolute support per class, integrative score w and percentage of testing observations (under 10-CV) with θ=0.8 matching threshold.

Clinical Patterns. Tables 10 and 11 display the top discriminative rules (according to their weighted confidence and lift) for the *breast*, *statlog*, *hungarian*, *hepatitis*, *postoperative*, *lung*, and *mammography* datasets. Each rule is characterized by the underlying pattern, $\varphi_{\mathbf{J}}$, and number of observations $|I|$ that strictly satisfy $\varphi_{\mathbf{J}}$. For this analysis, we applied FleBiC with default behavior. Weighted confidence should be assessed against the total number of classes,

connective		*fibro-adenoma*	
φ={k_8=2}	$\|I\|$=5 wlift=7.57 wconf=1.0	φ={k_0=0,k_1=0,k_4=1,k_6=0,k_8=0}	$\|I\|$=8 wlift=4.48 wconf=0.63
φ={k_1=1,k_4=2}	$\|I\|$=5 wlift=5.4 wconf=0.71	φ={k_4=1,k_5=1,k_7=1}	$\|I\|$=8 wlift=2.69 wconf=0.38
φ={k_2=1,k_3=3,k_5=1,k_6=2,k_7=3}	$\|I\|$=6 wlift=6.39 wconf=0.84	φ={k_0=0,k_3=0,k_4=1,k_6=0,k_8=0}	$\|I\|$=6 wlift=2.66 wconf=0.37
φ={k_4=1,k_6=1,k_7=2}	$\|I\|$=5 wlift=4.73 wconf=0.62	φ={k_0=0,k_4=1,k_5=1,k_8=0}	$\|I\|$=6 wlift=2.46 wconf=0.34
φ={k_1=1,k_3=3,k_7=3}	$\|I\|$=5 wlift=3.78 wconf=0.5	φ={k_1=0,k_4=1}	$\|I\|$=7 wlift=1.97 wconf=0.28
φ={k_1=0,k_4=1,k_6=1}	$\|I\|$=6 wlift=3.49 wconf=0.46	φ={k_4=1,k_5=1}	$\|I\|$=10 wlift=1.85 wconf=0.26
carcinoma		*fibro-glandular*	
φ={k_0=1,k_1=3,k_2=3,k_8=1}	$\|I\|$=11 wlift=4.68 wconf=0.92	φ={k_2=1,k_3=0,k_4=1,k_7=0}	$\|I\|$=10 wlift=3.8 wconf=0.57
φ={k_2=3,k_8=1}	$\|I\|$=12 wlift=4.65 wconf=0.92	φ={k_3=0,k_4=1,k_6=1}	$\|I\|$=12 wlift=3.78 wconf=0.57
φ={k_1=3,k_4=1,k_8=1}	$\|I\|$=12 wlift=4.32 wconf=0.85	φ={k_0=0,k_2=1,k_3=0,k_4=1,k_7=0,k_8=0}	$\|I\|$=12 wlift=3.38 wconf=0.51
φ={k_1=3,k_6=1,k_8=1}	$\|I\|$=10 wlift=4.2 wconf=0.83	φ={k_1=2,k_4=1}	$\|I\|$=9 wlift=3.13 wconf=0.47
φ={k_1=3,k_3=1,k_7=1,k_8=1}	$\|I\|$=13 wlift=2.88 wconf=0.57	φ={k_3=0,k_4=1}	$\|I\|$=15 wlift=3.1 wconf=0.46
φ={k_0=1,k_8=1}	$\|I\|$=19 wlift=2.52 wconf=0.5	φ={k_0=0,k_3=0,k_4=1,k_6=1,k_8=0}	$\|I\|$=9 wlift=2.88 wconf=0.43
mastopathy		*adipose*	
φ={k_0=0,k_4=1,k_6=1}	$\|I\|$=8 wlift=2.14 wconf=0.36	φ={k_0=3,k_3=3,k_5=3,k_6=3,k_7=3,k_8=3}	$\|I\|$=14 wlift=4.81 wconf=1.0
φ={k_3=1,k_4=1,k_6=1,k_7=1,k_8=1}	$\|I\|$=9 wlift=2.02 wconf=0.34	φ={k_0=3,k_1=0,k_6=3,k_8=3}	$\|I\|$=7 wlift=4.2 wconf=0.87
φ={k_4=1,k_5=1,k_6=1}	$\|I\|$=10 wlift=1.89 wconf=0.32	φ={k_0=3,k_1=0,k_8=3}	$\|I\|$=14 wlift=3.96 wconf=0.82
φ={k_2=0,k_4=1}	$\|I\|$=8 wlift=1.88 wconf=0.32	φ={k_0=3,k_5=3,k_7=3,k_8=3}	$\|I\|$=11 wlift=3.93 wconf=0.81
φ={k_0=1,k_3=1,k_4=1,k_7=1,k_8=1}	$\|I\|$=7 wlift=1.88 wconf=0.32	φ={k_0=3,k_8=3}	$\|I\|$=22 wlift=3.92 wconf=0.81
φ={k_0=0,k_4=1}	$\|I\|$=11 wlift=1.79 wconf=0.3	φ={k_0=3,k_1=0,k_4=2,k_8=3}	$\|I\|$=9 wlift=3.9 wconf=0.81

Table 10: Discriminative patterns of small length for the six classes of the *breast* tissue dataset (under constant and additive coherence assumptions). The *breast* variables are numeric where y_0=impedivity at 0Hz, y_1=phase angle at 500KHz, y_2=slope of phase angle, y_3=distance between spectral ends, y_4=area under spectrum, y_5=normalized area, y_6=maximum, y_7=distance to maximum, y_8=length of spectral curve. To simplify the presentation, the real values within φ patterns were mapped into an ordinal 0-3 scale. Illustrating, the *additive* pattern {k_2=3,k_8=1} indicates that phase angles with accentuated slope ($y_2 \in$\{2,3\}) and shorter length of spectral curve ($y_8 \in$\{0,1\}) are likely to be associated with the carcinoma cancer type.

$1/|\mathcal{C}|$ (e.g. 1/3 for *lung* data and 1/6 for *breast* data). Regarding weighted lift, 1 is the non-discriminative value reference (pattern-class independence). Results suggest the relevance of the found patterns. Non-constant patterns were found for datasets with real-valued variables (e.g. breast dataset), further underlining their role to discriminate clinical conditions (Figure 8).

Interestingly, the found patterns per dataset can radically differ regarding support and length, motivating the importance of being able to flexibly select subspaces when learning from high-dimensional data. Finally, the discovered patterns are interpretable and clinically meaningful.

6. Conclusions and Future Work

This work addressed the problem of classifying (high-dimensional) biomedical data. Motivated by the well-recognized relevance of non-constant patterns and limitations of existing associative classifiers, we proposed a new classifier, FleBiC. Aided by state-of-the-art contributions on pattern-based biclustering, FleBiC uses discriminative patterns with diverse forms of coherence (including constant, additive, multiplicative and order-preserving assumptions) and quality to learn from high-dimensional data.

FleBiC holds singular properties of interest: *1)* discover coherent, statistically significant, and discriminative patterns; *2)* place adequate scoring criteria to tolerate noise, highlight relevant patterns and weight different co-

statlog: control		*statlog*: heart disease	
φ={k_8=0,k_{11}=0,k_{12}=0}	$\|I\|$=82 wlift=1.62 wconf=0.9	φ={k_2=3,k_{12}=3}	$\|I\|$=68 wlift=1.96 wconf=0.87
φ={k_5=1,k_{11}=0,k_{12}=0}	$\|I\|$=85 wlift=1.57 wconf=0.87	φ={k_{10}=3,k_{12}=3}	$\|I\|$=67 wlift=1.83 wconf=0.81
φ={k_{10}=0,k_{12}=0}	$\|I\|$=82 wlift=1.57 wconf=0.87	φ={k_1=3,k_2=3,k_5=1}	$\|I\|$=65 wlift=1.72 wconf=0.76
φ={k_5=1,k_{10}=0}	$\|I\|$=82 wlift=1.34 wconf=0.74	φ={k_8=3}	$\|I\|$=66 wlift=1.66 wconf=0.74
hungarian: <50% diameter narrowing		*hungarian*: >50% diameter narrowing	
φ={k_6=0,k_8=0,k_9=1}	$\|I\|$=124 wlift=1.34 wconf=0.86	φ={k_1=1,k_2=3,k_8=1}	$\|I\|$=56 wlift=2.5 wconf=0.9
φ={k_5=0,k_9=1}	$\|I\|$=147 wlift=1.31 wconf=0.84	φ={k_1=1,k_{10}=1}	$\|I\|$=62 wlift=2.42 wconf=0.87
φ={k_5=0,k_6=0,k_9=1}	$\|I\|$=120 wlift=1.3 wconf=0.83	φ={k_2=3,k_8=1}	$\|I\|$=61 wlift=2.34 wconf=0.84
mammography: benign		*mammography*: malign	
φ={k_0=4,k_1=0,k_3=2}	$\|I\|$=140 wlift=1.7 wconf=0.91	φ={k_0=5,k_1=3,k_3=2}	$\|I\|$=215 wlift=1.97 wconf=0.91
φ={k_0=4,k_2=0}	$\|I\|$=288 wlift=1.69 wconf=0.91	φ={k_0=5,k_2=3}	$\|I\|$=134 wlift=1.91 wconf=0.88
φ={k_1=0,k_2=0}	$\|I\|$=163 wlift=1.64 wconf=0.88	φ={k_1=3}	$\|I\|$=315 wlift=1.7 wconf=0.78
hepatitis: live (outcome)		*postoperative*: home care	
φ={k_1=0,k_{10}=0}	$\|I\|$=22 wlift=2.36 wconf=0.48	φ={k_1=2,k_3=2,k_4=0}	$\|I\|$=7 wlift=1.45 wconf=0.38
φ={k_1=0,k_4=0,k_5=0,k_{18}=1}	$\|I\|$=23 wlift=2.13 wconf=0.44	φ={k_1=2,k_4=0,k_7=3}	$\|I\|$=6 wlift=1.4 wconf=0.37
φ={k_1=0,k_7=1,k_{18}=1}	$\|I\|$=20 wlift=1.97 wconf=0.4	φ={k_2=1,k_6=1}	$\|I\|$=9 wlift=1.35 wconf=0.36
φ={k_1=0,k_3=1,k_4=0,k_{10}=0}	$\|I\|$=22 wlift=1.85 wconf=0.38	φ={k_3=2,k_6=1}	$\|I\|$=11 wlift=1.33 wconf=0.35
φ={k_1=0,k_2=0,k_4=0}	$\|I\|$=19 wlift=1.84 wconf=0.38	φ={k_3=2,k_4=1,k_5=1}	$\|I\|$=9 wlift=1.29 wconf=0.34
φ={k_1=0,k_4=0,k_6=1}	$\|I\|$=20 wlift=1.64 wconf=0.33	φ={k_1=2,k_3=2,k_6=1,k_7=3}	$\|I\|$=7 wlift=1.27 wconf=0.34

Table 11: Discriminative patterns of small length in the *statlog*, *hungarian*, *mammography*, *hepatitis* and *postoperative* datasets. Illustrating, the constant pattern {k_1=1,k_7=1,k_9=1,k_{11}=1,k_{12}=1, k_{13}=1} from *hepatitis* dataset indicates that women (y_1) with normal liver (y_7), absence of ascites (y_{11}) and varices (y_{12}) and low levels of bilirubin (y_{13}) are likely to survive. Order-preserving patterns from real-valued variables can be also found. Illustrating, {k_1=0,k_2=1,k_6=2} ($k_1 \leq k_2 < k_6$) from *postoperative* data indicates that individuals with high external temperature (low y_1), mid-to-scarce oxygen levels (y_2) and unstable blood pressure (higher y_6) are likely to require intensive care. The discriminative pattern φ={k_2=3,k_6=3,k_{11}=3,k_{12}=3} in *statlog* and *hungarian* data indicates that individuals with chest pain (y_2), electrocardiac complications (y_6), signaled vessels (y_{11}), and high thal (y_{12}) are likely to suffer from heart disease. FleBiC can thus properly diagnosis patients having a subset of all possible symptoms by properly weighting the matched patterns per class.
Attributes of *hepatitis* dataset are: y_0=age (0-7), y_1=sex, y_2=steroid, y_3=antivirals, y_4=fatigue, y_5=malaise, y_6=anorexia, y_7=liver big, y_8=liver firm, y_9=spleen palpable, y_{10}=spiders, y_{11}=ascites, y_{12}=varices, y_{13}=bilirubin (0-5 level), y_{14}=alk phosphate (0-5), y_{15}=sgot (0-5), y_{16}=albumin (0-5), y_{17}=protime (0-8), y_{18}=histology (where y_2 to y_{12} and y_{18} are binary). The *postoperative* variables are y_0=IT internal temperature (0-3 where 0 is high), y_1=ST surface temperature (0-3), y_2=oxygen saturation (0-3 where 0 is excellent), y_3=BP blood pressure (0-2 where 0 is high), y_4=ST stability (0-2 where 0 is stable), y_5=IT stability (0-2), y_6=BP stability (0-2) and y_7=perceived comfort (0-3). *statlog* and *hungarian* variables are y_0=age (0-3 where 0 is young), y_1=sex, y_2=chest pain type (0-3), y_3=resting blood pressure (0-3 where 0 is low), y_4=serum cholesterol (0-3), y_5=fasting blood sugar (1:high, 0:low), y_6=electrocardiogram (0:normal, 1:abnormal, 2:definite), y_7=maximum heart rate, y_8=induced angina (0-3), y_9=ST depression (yes/no), y_{10}=slope of ST peak (0-3), y_{11}=number of vessels colored by flourosopy (0-3), y_{12}=thal (0-2).

herencies that prevent some patterns from jeopardizing the learning; and *3)* apply robust matching criteria that verify shift, scaling and order-preserving factors, and accommodate noise penalizations. FleBiC can efficiently generate (and score) association rules with disjunctions of labels in the consequent to further learn from patterns able to discriminate more than a single class. Finally, FleBiC is able to combine the output of global learning functions for datasets meaningfully described by both local and global regularities.

Results on biological, clinical and synthetic data confirm the underly-

ing hypothesis of our work: combining discriminative patterns with varying coherence and quality improves associative classification. Comparison with state-of-the-art classifiers show the relevance of learning from non-constant patterns. In addition to these observations, the inherent interpretability of the learned associative classification models turn them state-of-the-art candidates to describe phenotypes and support medical decisions.

This work opens a critical door to understand how subspaces with varying coherence and quality impact descriptive and predictive tasks across biomedical domains. The possibility to parameterize the desirable properties of the targeted patterns can be further considered to systematically study the intricacies of complex regulatory behavior in omic data and physiological responses in clinical data. Furthermore, the gathered observations can be used to extend state-of-the-art (non-associative) classifiers, revising their behavior to model non-constant relationships between data observations.

Acknowledgments

This work was supported by *Fundação para a Ciência e a Tecnologia* under projects ILU (DSAIPA/DS/0111/2018), NEUROCLINOMICS2 (PTDC/EEI-SII/1937/2014), PREDICT (PTDC/CCI-CIF/29877/2017) and AIpALS (PTDC/CCI-CIF/4613/2020), and INESC-ID (UIDB/50021/2020) and LASIGE (UIDB/00408/2020, UIDP/00408/2020) pluriannuals.

References

1. P. Bühlmann, S. Van De Geer, Statistics for high-dimensional data: methods, theory and applications, Springer Science & B. Media, 2011.
2. R. Henriques, F. L. Ferreira, S. C. Madeira, Bicpams: software for biological data analysis with pattern-based biclustering, BMC bioinformatics 18 (1) (2017) 82.
3. R. Henriques, S. C. Madeira, Towards robust performance guarantees for models learned from high-dimensional data, in: Big Data in Complex Systems, Springer, 2015, pp. 71–104.
4. V. N. Vapnik, Statistical Learning Theory, Wiley-Interscience, 1998.
5. M. A. Figueiredo, A. K. Jain, M. H. Law, A feature selection wrapper for mixtures, in: Pattern Recognition and Image Analysis, Springer, 2003, pp. 229–237.
6. N. Cristianini, J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods, Cambridge university press, 2000.
7. M. A. Figueiredo, A. 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, Vol. 1, IEEE, 2001, pp. I–35.
8. M. Wang, X. Shang, X. Li, Z. Li, W. Liu, Efficient mining differential co-expression constant row bicluster in real-valued gene expression datasets, Applied Mathematics & Information Sciences 7 (2) (2013).
9. L. Parsons, E. Haque, H. Liu, Subspace clustering for high dimensional data: A review, SIGKDD Exp. Newsl. 6 (1) (2004) 90–105.
10. R. Henriques, Learning from high-dimensional data using local descriptive models, Ph.D. thesis, Instituto Superior Tecnico, Universidade de Lisboa, Lisboa (2016).
11. M. Almasi, M. S. Abadeh, Cars-lands: An associative classifier for large-scale datasets, Pattern Recognition 100 (2020) 107128.
12. B. Bringmann, S. Nijssen, A. Zimmermann, Pattern-based classification: a unifying perspective, arXiv:1111.6191 (2011).

13. O. Odibat, C. K. Reddy, Efficient mining of discriminative co-clusters from gene expression data, KAIS (2013) 1–30.
14. G. Nayak, V. Mithal, X. Jia, V. Kumar, Classifying multivariate time series by learning sequence-level discriminative patterns, in: Proceedings of the 2018 SIAM International Conference on Data Mining, SIAM, 2018, pp. 252–260.
15. R. Henriques, S. Madeira, Bicpam: Pattern-based biclustering for biomedical data analysis, Algorithms for Molecular Biology 9 (1) (2014) 27.
16. R. Henriques, C. Antunes, S. C. Madeira, Generative modeling of repositories of health records for predictive tasks, Data Mining and Knowledge Discovery 29 (4) (2015) 999–1032.
17. A. V. Carreiro, O. Anunciação, J. A. Carriço, S. C. Madeira, Prognostic prediction through biclustering-based classification of clinical gene expression time series., Journal of integrative bioinformatics 8 (3) (2010) 175–175.
18. R. Henriques, S. Madeira, Biclustering with flexible plaid models to unravel interactions between biological processes, IEEE/ACM Trans. Comput. Biol. Bioinformatics (2015).
19. S. Hochreiter, U. Bodenhofer, M. Heusel, A. Mayr, A. Mitterecker, A. Kasim, T. Khamiakova, S. Van Sanden, D. Lin, W. Talloen, L. Bijnens, H. W. H. Göhlmann, Z. Shkedy, D.-A. Clevert, FABIA: factor analysis for bicluster acquisition, Bioinformatics 26 (12) (2010) 1520–1527.
20. M. Alzahrani, H. Kuwahara, W. Wang, X. Gao, Gracob: a novel graph-based constant-column biclustering method for mining growth phenotype data, Bioinformatics (2017) btx199.
21. R. Henriques, S. C. Madeira, Bicnet: Flexible module discovery in large-scale biological networks using biclustering, Algorithms for Molecular Biology 11 (1) (2016) 1–30.
22. C. Ding, Y. Zhang, T. Li, S. R. Holbrook, Biclustering protein complex interactions with a biclique finding algorithm, in: ICDM, IEEE Computer Society, 2006, pp. 178–187.
23. S. Wang, R. R. Gutell, D. P. Miranker, Biclustering as a method for rna local multiple sequence alignment, Bioinformatics 23 (24) (2007) 3289–3296.
24. J. Liu, W. Wang, Op-cluster: Clustering by tendency in high dimensional space, in: ICDM, IEEE CS, 2003, pp. 187–.
25. L. Lazzeroni, A. Owen, Plaid models for gene expression data, Statistica Sinica 12 (2002) 61–86.
26. S. C. Madeira, A. L. Oliveira, Biclustering algorithms for biological data analysis: A survey, IEEE/ACM Trans. Comput. Biol. Bioinformatics 1 (1) (2004) 24–45.
27. A. Ben-Dor, B. Chor, R. Karp, Z. Yakhini, Discovering local structure in gene expression data: the order-preserving submatrix problem, in: RECOMB, ACM, 2002, pp. 49–57.
28. R. Henriques, S. Madeira, Bicspam: Flexible biclustering using sequential patterns, BMC Bioinformatics 15 (2014) 130.
29. W. Li, J. Han, J. Pei, Cmar: Accurate and efficient classification based on multiple class-association rules, in: ICDM, IEEE Computer Society, 2001, pp. 369–376.
30. R. Ramírez-Rubio, M. Aldape-Pérez, C. Yáñez-Márquez, I. López-Yáñez, O. Camacho-Nieto, Pattern classification using smallest normalized difference associative memory, Pattern Recognition Letters 93 (2017) 104–112.
31. T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM, 2016, pp. 785–794.
32. A. P. White, W. Z. Liu, Bias in information-based measures in decision tree induction, Machine Learning 15 (3) (1994) 321–329.
33. J. Ma, Y. Zhang, L. Zhang, Discriminative subspace matrix factorization for multiview data clustering, Pattern Recognition 111 107676.
34. J. P. Goncalves, Integrative mining of gene regulation and its perturbations, Ph.D. thesis, Instituto Superior Tecnico, Universidade de Lisboa, Lisboa (2013).
35. N. Lesh, M. J. Zaki, M. Ogihara, Mining features for sequence classification, in: KDD, ACM, 1999, pp. 342–346.
36. P. Geurts, Pattern extraction for time series classification, in: Principles of Data Mining and Knowl. Disc., Vol. 2168 of LNCS, Springer, 2001, pp. 115–127.
37. J. Shang, W. Tong, J. Peng, J. Han, Dpclass: An effective but concise discriminative patterns-based classification framework, in: Proceedings of the 2016 SIAM International Conference on Data Mining, SIAM, 2016, pp. 567–575.

38. I. Tagkopoulos, N. Slavov, S.-Y. Kung, Multi-class biclustering and classification based on modeling of gene regulatory networks, in: BIBE, IEEE, 2005, pp. 89–96.
39. V. Tseng, C.-H. Lee, Effective temporal data classification by integrating sequential pattern mining and probabilistic induction, Expert Sys.App. 36 (5) (2009) 9524–9532.
40. T. P. Exarchos, M. G. Tsipouras, C. Papaloukas, D. I. Fotiadis, A two-stage methodology for sequence classification based on sequential pattern mining and optimization, Data Knowl. Eng. 66 (3) (2008) 467–487.
41. E. Loekito, J. Bailey, Using highly expressive contrast patterns for classification-is it worthwhile?, in: Advances in Know. Disc. and Data Mining, Springer, 2009, pp. 483–490.
42. R. Henriques, C. Antunes, Learning predictive models from integrated healthcare data: Extending pattern-based and generative models to capture temporal and cross-attribute dependencies, in: HICSS, 2014, pp. 2562–2569.
43. A. Veloso, J. Wagner Meira, M. J. Zaki, Lazy associative classification, in: ICDM, IEEE, 2006.
44. B. Liu, W. Hsu, Y. Ma, Integrating classification and association rule mining, in: KDD, ACM, 1998, pp. 80–86.
45. G. Dong, J. Li, Efficient mining of emerging patterns: Discovering trends and differences, in: KDD, ACM, 1999, pp. 43–52.
46. X. Yin, J. Han, Cpar: Classification based on predictive association rules, in: SDM, Vol. 3, SIAM, 2003, pp. 331–335.
47. G. Cong, K.-L. Tan, A. K. H. Tung, X. Xu, Mining top-k covering rule groups for gene expression data, in: SIGMOD, ACM, 2005, pp. 670–681.
48. J. R. Quinlan, Induction of decision trees, Machine learning 1 (1) (1986) 81–106.
49. J. Wang, G. Karypis, Harmony: Efficiently mining the best rules for classification, in: SDM, Vol. 5, SIAM, 2005, pp. 205–216.
50. H. Cheng, X. Yan, J. Han, P. S. Yu, Direct discriminative pattern mining for effective classification, in: ICDE, IEEE, 2008, pp. 169–178.
51. W. Fan, K. Zhang, H. Cheng, J. Gao, X. Yan, J. Han, P. Yu, O. Verscheure, Direct mining of discriminative and essential frequent patterns via model-based search tree, in: KDD, ACM, 2008, pp. 230–238.
52. A. Zimmermann, B. Bringmann, Aggregated subset mining, in: Advances in Knowledge Discovery and Data Mining, Springer, 2009, pp. 664–672.
53. D. Nielsen, Tree boosting with xgboost-why does xgboost win" every" machine learning competition?, Master's thesis, NTNU (2016).
54. A. Tanay, R. Sharan, R. Shamir, Discovering statistically significant biclusters in gene expression data, Bioinformatics 18 (2002) 136–144.
55. C. Tang, L. Zhang, M. Ramanathan, A. Zhang, Interrelated two-way clustering: An unsupervised approach for gene expression data analysis, in: BIBE, IEEE CS, 2001, pp. 41–.
56. J. A. Hartigan, Direct Clustering of a Data Matrix, Journal of the American Statistical Association 67 (337) (1972) 123–129.
57. M. Wang, X. Shang, S. Zhang, Z. Li, Fdcluster: mining frequent closed discriminative bicluster without candidate maintenance in multiple microarray datasets, in: Data Mining Workshops (ICDMW), IEEE, 2010, pp. 779–786.
58. Y. Cheng, G. M. Church, Biclustering of expression data, in: Int. Sys. for Molec. Biology, AAAI, 2000, pp. 93–103.
59. J. Yang, W. Wang, H. Wang, P. Yu, delta-clusters: capturing subspace correlation in a large data set, in: ICDE, 2002, pp. 517 –528.
60. X. Gan, A. Liew, H. Yan, Discovering biclusters in gene expression data based on high-dimensional linear geometries, BMC Bioinformatics 9 (1) (2008) 209.
61. B. Pontes, R. Giráldez, J. S. Aguilar-Ruiz, Configurable pattern-based evolutionary biclustering of gene expression data, Algorithms for Molecular Biology 8 (1) (2013) 4.
62. F. de Franga, F. Von Zuben, Extracting additive and multiplicative coherent biclusters with swarm intelligence, in: Evolutionary Computation (CEC), 2011, pp. 632–638.
63. Q. Fang, W. Ng, J. Feng, Y. Li, Mining order-preserving submatrices from probabilistic matrices, ACM Trans. Database Syst. 39 (1) (2014) 6:1–6:43.
64. R. Henriques, C. Antunes, S. C. Madeira, A structured view on pattern mining-based biclustering, Pattern Recognition 4 (12) (2015) 3941—-3958.
65. R. Henriques, S. C. Madeira, Bic2pam: constraint-guided biclustering for biological data

analysis with domain knowledge, Algorithms for Molecular Biology 11 (1) (2016) 23.

66. Y. Okada, W. Fujibuchi, P. Horton, A biclustering method for gene expression module discovery using closed itemset enumeration algorithm, IPSJ Trans. on Bioinf. 48 (SIG5) (2007) 39–48.
67. A. Serin, M. Vingron, Debi: Discovering differentially expressed biclusters using a frequent itemset approach, Algorithms for Molecular Biology 6 (2011) 1–12.
68. R. Henriques, S. C. Madeira, Bsig: evaluating the statistical significance of biclustering solutions, Data Mining and Knowledge Discovery 32 (1) (2018) 124–161.
69. L. Zhang, S. K. Shah, I. A. Kakadiaris, Hierarchical multi-label classification using fully associative ensemble learning, Pattern Recognition 70 (2017) 89–103.
70. A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, et al., Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling, Nature 403 (6769) (2000) 503–511.
71. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, science 286 (5439) (1999) 531–537.
72. S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. Kim, L. C. Goumnerova, P. M. Black, C. Lau, et al., Prediction of central nervous system embryonal tumour outcome based on gene expression, Nature 415 (6870) (2002) 436–442.
73. U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, A. J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proceedings of the National Academy of Sciences 96 (12) (1999) 6745–6750.
74. M. Lichman, UCI machine learning repository (2013).
URL `http://archive.ics.uci.edu/ml`
75. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The weka data mining software: An update, SIGKDD Explor. 11 (1) (2009) 10–18.