Knowl Inf Syst (2012) 30:87–111
DOI 10.1007/s10115-010-0369-x

REGULAR PAPER

Application-independent feature construction based on almost-closedness properties

Dominique Gay · Nazha Selmaoui-Folcher · Jean-François Boulicaut

Received: 8 July 2009 / Revised: 20 October 2010 / Accepted: 26 November 2010 /
Published online: 25 December 2010

Abstract Feature construction has been studied extensively, including for 0/1 data samples. Given the recent breakthroughs in closedness-related constraint-based mining, we are considering its impact on feature construction for classification tasks. We investigate the use of condensed representations of frequent itemsets based on closedness properties as new features. These itemset types have been proposed to avoid set counting in difficult association rule mining tasks, i.e. when data are noisy and/or highly correlated. However, our guess is that their intrinsic properties (say the maximality for the closed itemsets and the minimality for the δ-free itemsets) should have an impact on feature quality. Understanding this remains fairly open, and we discuss these issues thanks to itemset properties on the one hand and an experimental validation on various data sets (possibly noisy) on the other hand.

Keywords Feature construction · Pattern-based classification · δ-free itemsets · δ-strong rules · Closure equivalence classes · Noise-tolerance

1 Introduction

Feature construction is one of the major research topics for supporting classification tasks. Based on a set of original features, the idea is to compute new features that may better describe

Dominique Gay was with PPME EA3325, ERIM EA3791, University of New-Caledonia when this work began.

D. Gay (✉) · N. Selmaoui-Folcher
University of New-Caledonia, PPME EA3325, ERIM EA3791,
BP R4, 98851 NOUMEA Cédex, New-Caledonia, France
e-mail: dominique.gay@univ-nc.nc

D. Gay
Orange Labs, TECH/ASAP/PROF, 2, avenue Pierre Marzin, 22307 LANNION Cédex, France
e-mail: dominique.gay@orange-ftgroup.com

J.-F. Boulicaut
INSA-Lyon, LIRIS CNRS UMR5205, INRIA Combining, 69621 Villeurbanne, France

labeled samples as the predictive accuracy of classifiers can be improved. First approaches have focused on extending classical univariate decision trees to multivariate decision trees: using linear combinations of original attributes as new features [13,51] or X-of-N attributes for splitting the data [61]. More generally, when not restricting to decision tree learning, several authors (e.g., [17]) have proposed to look at feature construction based on frequent patterns (i.e., in most of the cases, collections of attribute-value pairs which are true or not within a sample)—keeping in mind that frequent itemsets could bring more information than single items.

Using patterns (e.g., itemsets or association rules) that hold in 0/1 data for classification purpose is not new. Indeed, the pioneering work on pattern-based classification and the CBA method [40]) has given rise to many proposals: [2,20,39] and [21,22,36,37] ; see also [11] and [46] for recent surveys. Even if various pattern-based approaches exist, most of them share the same main concepts: patterns should be interesting w.r.t. a given class-discriminant interestingness measure, the set of patterns should cover *well* the training data set while being concise, and minimal non-redundant patterns should be preferred.

Recent breakthroughs in frequent pattern mining deal with redundancy and conciseness by means of condensed representations (i.e., rather small collections of patterns from which one can infer the frequency of many sets instead of counting for it [14]). Recent feature construction approaches benefit from such advances: in [17,27,38,48], the authors focus on condensed representations based on support equivalence classes (also called closure equivalence classes), i.e., groups of (frequent) itemsets (sets of Boolean attributes) supporting the same sets of objects. The ones bet on the maximal elements of the equivalence classes, the so-called closed itemsets. The others explain that we must use their minimal elements that are also called generators [4] or 0-free itemsets [8].

In this paper, we clarify this divergence of arguments, and we use extensions of such concepts to contribute to difficult classification tasks. Indeed, we investigate such a feature construction in the presence of attribute noise. In noisy dense binary data, looking for free/closed itemsets is known to be computationally hard. Furthermore, the relevancy of such patterns as features that describe well the underlying data is questionable. In [9], a ”near equivalence” perspective has been proposed. It has then be exploited in various settings like fault-tolerant pattern mining [5] or cluster characterization in real-life data [44]. Our guess is that it could also be useful in a noise-tolerant feature construction process. Our main contribution is then methodological. We discuss the intrinsic properties of the various pattern types when targeting a classification process. We formalize the definition of ”almost-closure equivalence classes”, and we show how to use it during a feature construction process. The two main steps of the process consist in:

- the efficient extraction of pattern sets in which patterns satisfy application-independent constraints,
- the encoding of the original data into a new data set by using extracted patterns as new features.

Then, we can use the transformed data sets to learn classifiers thanks to available methods. To support our empirical study, we use different tools, namely implementations of C4.5 [45], Naive Bayes [34], and support vector machines. We also compare with one of the best pattern-based proposal, namely HARMONY [53,54]. In our previous work [30,31], we introduced such a generic feature construction, and we started to study attribute noise-tolerance. This paper provides much more details about the whole process. Design decisions are discussed in depth, and we provide formal arguments. Last but not the least, the whole experimental validation has been considerably extended.

The rest of the paper is organized as follows. We start by setting the context of feature construction based on patterns, and we provide the needed preliminary definitions in Sect. 2. In Sect. 3, we start from a comparison between two existing approaches using closedness properties to motivate our proposal. Then, in Sect. 4, we introduce class-discriminant δ-closure equivalence classes. Their properties are discussed in Sect. 5. Our application-independent feature construction process is defined in Sect. 6, and it is empirically evaluated in Sect. 7. Section 8 deals with the limits of our method, and it also comments the related work that has not been considered so far. Finally, Sect. 9 concludes.

2 Context and preliminary definitions

We consider feature construction when data samples correspond to 0/1 data. Let us first recall some standard terminology in that context. A binary database $r = \{\mathcal{T}, \mathcal{I}, \mathcal{R}\}$ is built on $\mathcal{T}$ which is a set of objects (or transactions) described by a set $\mathcal{I}$ of Boolean attributes (or items). We have $\mathcal{R} : \mathcal{T} \times \mathcal{I} \mapsto \{0, 1\}$. When $\mathcal{R}(t, i) = 1$, we say that object t satisfies attribute i (or that transaction t contains item i). An itemset $I \subseteq \mathcal{I}$ is a set of items. The frequency of an itemset $I \subseteq \mathcal{I}$ is $freq(I, r) = |Objects(I, r)|$ where $Objects(I, r) = \{t \in \mathcal{T} \mid \forall i \in \mathcal{I} \quad \mathcal{R}(t, i) = 1\}$. $Objects(I, r)$ are also called the *support* of I. An itemset I is said to be γ-frequent if $freq(I, r) \geq \gamma$. We now motivate the study of closedness-related properties when considered the use of frequent itemsets for class characterization.

2.1 Interestingness of frequent itemsets

Considering that *"a frequent itemset might be interesting"* is intuitive. In [17], the authors have linked the frequency with other interestingness measures such as Information Gain and Fischer score. They also explain why frequent patterns should be used for feature construction. Notice however that for a very low frequency threshold γ, frequent itemsets are numerous, but only few of them are interesting. Conversely, for a very high frequency threshold, interesting itemsets are becoming rare, and it can be difficult to obtain a nice coverage of the training data. Since even for valuable frequency thresholds, the number of frequent itemsets can be huge in dense databases [6], and given that the whole collection of frequent itemsets is generally hard to compute, it is now common to use condensed representations to save space and time during such a pattern mining task (see, e.g., [14] for a survey paper). A condensed representation for frequent itemsets can be seen as a subset of the whole collection of frequent itemsets from which one can derive every missing frequent itemset and infer its frequency.

2.2 Equivalence classes of itemsets

Definition 1 (*Closed itemset*) An itemset I is a *closed itemset* in r iff there is no superset of I with the same frequency as I in r, i.e., $\nexists I' \supset I$ s.t. $freq(I', r) = freq(I, r)$.

Another definition exploits the closure operator $cl : \mathcal{P}(\mathcal{I}) \rightarrow \mathcal{P}(\mathcal{I})$. Let us assume that $Items$ is the dual operator for $Objects$: given $T \subseteq \mathcal{T}$, $Items(T, r) = \{i \in \mathcal{I} \mid \forall t \in T, (t, i) \in R\}$. We now define $cl(I, r) \equiv Items(Objects(I, r), r)$. By definition, an itemset I is a closed itemset in r iff $I = cl(I, r)$.

This definition of the closure operator cl has been extensively used in the context of *Formal Concept Analysis* [26]. Since [4], it is common to formalize the fact that many itemsets have the same closure, and thus the same frequency, by means of *closure equivalence* classes.

Table 1 A toy example of a binary-labeled database

r	A	B	C	D	c_1	c_2
t_1	1	1	1	1	1	0
t_2	1	1	1	0	1	0
t_3	0	1	1	0	1	0
t_4	1	0	0	1	1	0
t_5	0	1	1	0	0	1
t_6	0	1	0	1	0	1
t_7	1	0	1	1	0	1

Definition 2 (*Closure equivalence, free itemset*) Two itemsets I and J are said to be *equivalent* in r (denoted $I \sim_{cl} J$) iff $cl(I, r) = cl(J, r)$. Thus, a *closure equivalence class* (CEC) is made of itemsets that have the same closure, i.e., they are all supported by the same set of objects ($Objects(I, r) = Objects(J, r)$). That is why, they are also known as support equivalence classes.

Each CEC contains exactly one maximal itemset (w.r.t. set inclusion) that is a closed itemset. It may contain several minimal itemsets that called key patterns in [4] and 0-free itemsets in [8]. More formally, an itemset I is a *free itemset* in r iff there is no subset of I with the same support (or the same closure) as I, i.e., $\nexists I' \subset I$ s.t. $freq(I', r) = freq(I, r)$ (or $cl(I', r) = cl(I, r)$).

Example 1 Considering Table 1, we have $r = (\mathcal{T}, \mathcal{I}, R)$, $\mathcal{T} = \{t_1, \ldots, t_7\}$, and $\mathcal{I} = \{A, B, C, D, c_1, c_2\}$, c_1 and c_2 being the class labels. For a frequency threshold $\gamma = 2$, itemsets AB and AC are γ-frequent. $ABCc_1$ is a γ-frequent closed itemset.

Considering the equivalence class $\mathcal{C} = \{AB, AC, ABC, ABc_1, ACc_1, ABCc_1\}$, AB and AC are its minimal elements (i.e., they are 0-free itemsets), and $ABCc_1$ is the maximal element, i.e., one of the closed itemsets in this toy database.

Since every itemset in a CEC has the same support, knowing the support of one itemset in a CEC enables to infer the support of the others. In other words, each element in a CEC carries the same information about the frequency as the other elements. For further details, we refer to the survey paper [14] but also to recent developments of such concepts (see, e.g., [7,41]). Therefore, if we need the frequency information, we may choose to keep only one element per CEC. Interestingly, when considering feature construction for binary-labeled databases, we find in the literature arguments for either closed itemsets or free itemsets.

3 Using freeness or closedness

Two different approaches for feature construction based on condensed representations have been considered so far. In, e.g., [38], the authors mine free itemsets (free itemsets are formally the 0-free itemsets) and closed itemsets (i.e., CECs) once the class attribute has been removed from the entire database. Other proposals, e.g., [17,27,28], consider (closed) itemset mining from samples of each class separately. In this section, the pros and cons of these alternatives are discussed.

Looking at the first direction of research, we may consider that closed sets, because of their maximality, are good candidates for characterizing labeled data, but not necessarily suitable to predict classes for unseen samples. Moreover, thanks to their minimality, free itemsets might be better for predictive tasks. Due to closedness properties, every itemset of a given closure equivalence class $\mathcal{C}$ in r covers exactly the same set of objects. Thus, free itemsets and their associated closed set are equivalent w.r.t. interestingness measures based on frequencies. As a result, to characterize a class, the choice between a free itemset and its closure remains open.

Let us now consider an incoming sample x (test phase) that is exactly described by the itemset Y (i.e., all its properties that are true are in Y). Furthermore, assume that we have $F \subseteq Y \subseteq cl(F, r)$ where F is a free itemset from the closure equivalence class $\mathcal{C}_F$. Using free itemsets to label x will not lead to the same decision than using closed itemsets. Indeed, $x \supseteq F$ and it satisfies Rule $F \to c$ while $x \not\supseteq cl(Y, r)$ and it does not satisfy Rule $cl(F, r) \to c$. Following that direction of work, in [3], authors have proposed classification rules whose bodies are free itemsets.

For the "per-class" approach, let us consider without loss of generality a two-class classification problem. In such a context, the equivalence between free itemsets and their associated closed ones is lost. The intuition is that, for a given free itemset Y in r_{c_1}—database restricted to samples of class c_1–and its closure $X = cl(Y, r_{c_1})$, X is more relevant than Y since $Objects(X, r_{c_1}) = Objects(Y, r_{c_1})$ and $Objects(X, r_{c_2}) \subseteq Objects(Y, r_{c_2})$. The closed itemsets (say $X = cl(X, r_{c_1})$) such that there is no other closed itemset (say $X' = cl(X', r)$) for which $cl(X, r_{c_2}) = cl(X', r_{c_2})$ are chosen as relevant itemsets to characterize c_1. In some cases, a free itemset Y can be equivalent to its closure $X = cl(Y, r_{c_1})$, i.e., $Objects(X, r_{c_2}) = Objects(Y, r_{c_2})$. Here, for the same reason as above, a free itemset may be chosen instead of its closed counterpart. Note that relevancy of closed itemsets does not avoid conflicting rules, i.e., we can have two closed itemsets: X relevant for c_1 and Y relevant for c_2 with $X \subseteq Y$.

Thus, choosing between freeness and closedness should be mainly motivated by the way data mining is performed on the entire data set or per class. Moreover, these two approaches do not take into account the class distribution, and both need for a post-processing to select discriminative patterns among the computed ones. In [12], several selection techniques are suggested to severely reduce the number of extracted patterns while keeping essential information. Here, we do not only look for closedness-related properties, but we also want to exploit interesting measures to keep only the disciminative ones. To avoid post-processing, we propose to use a syntactic constraint (i.e., keeping the class attribute during the mining phase) to mine class-discriminant closure equivalence classes.

4 Class-discriminant closure equivalence classes

A key property of CECs is that we can derive association rules [1] from them. We consider the association rules that can be derived from the itemsets of a same CEC. We also highlight the type of CEC that enables to derive class-discriminant rules.

Definition 3 (*Association rule*) An *association rule* π on r is an expression $I \to J$, where $I \subseteq \mathcal{I}$ and $J \subseteq \mathcal{I} \setminus I$. The *frequency* of the rule π is $freq(I \cup J, r)$ and its *confidence* is $conf(\pi, r) = freq(I \cup J, r)/freq(I, r)$. When $conf(\pi, r) = 1$, the rule π is said to be strong.

Example 2 From Table 1 and Example 1, we see that AB and ABC belong to the same CEC $\mathcal{C}$, and thus they have the same support. To consider the association rule $\pi : AB \to C$, we have $conf(\pi, r) = 1$. $AB \to C$ is a strong rule.

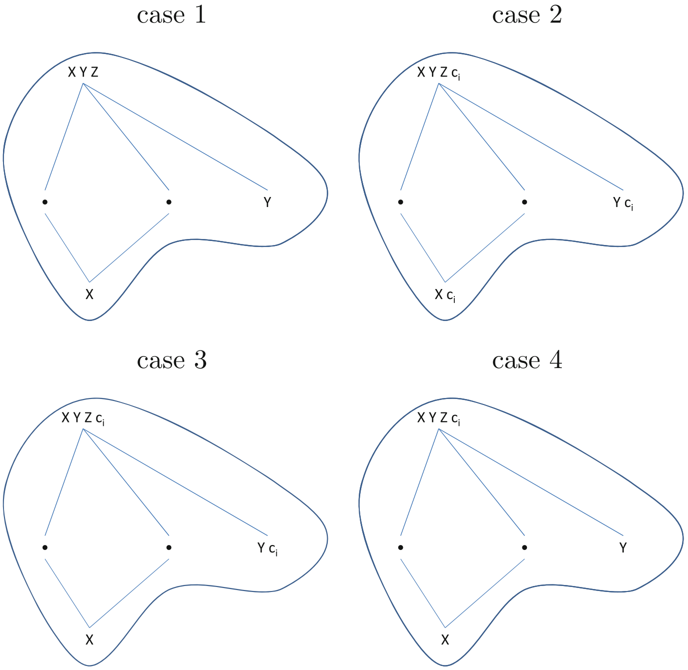

Fig. 1 Different types of CECs

When a CEC contains a class attribute, we can derive strong association rules that involve this class attribute. Indeed, given two itemsets X and Y from the same CEC such that $X \subseteq Y$, $X \rightarrow Y \setminus X$ is a strong association rule.

4.1 Mining with the class attribute

Keeping the class attribute when mining CECs may produce four different typical cases of CECs. Their differences depend on the way the class attribute appears or not in one or more itemsets of the CEC (see Fig. 1):

- In Case 1, the class attribute is not involved, so we cannot derive any association rule involving the class attribute.
- A CEC from Case 2 involves a class attribute. Unfortunately, all itemsets in the CEC contain it. Thus, we cannot derive any valuable association rule concluding on a class attribute.
- Cases 3 and 4 are interesting and quite similar. One (or more) free itemset does not contain the class attribute while the closed itemset contains it. In these cases, strong association rules concluding on a class attribute can be derived; here, e.g., $X \rightarrow c_i$.

Thus, interesting types of CECs correspond to Cases 3 and 4. Note also that in such cases, minimality and non-redundancy prevail: for example, in Case 3, we prefer X rather than any of its superset X' in the CEC to characterize class c_i. For the same reasons as in the previous section, X' appears redundant w.r.t. X.

In [49], the authors used these types of CECs to build a condensed representation of jumping emerging patterns (JEPs)—those patterns that appear only in one class of objects.

Clearly, JEPs are bodies of strong rules. JEPs have already been studied for classification purpose [37]. However, if the data are altered by noise, as it is usual in real-life applications, and given a reasonable frequency threshold γ, γ-frequent JEPs become rare and the collection of JEPs may not cover the database. To overcome this drawback, we propose to relax the support equivalence constraint in CECs.

4.2 From closures to δ-closures

To introduce softness in CEC and strong rule concepts, we suggest to allow some approximations (errors) when grouping itemsets. An integer parameter δ is used to specify the acceptable tolerance.

Definition 4 (*δ-strong rule, δ-free itemset*) Let δ be an integer. A *δ-strong rule* is an association rule of the form $I \rightarrow^{\delta} J$ which is violated in at most δ objects, and where $I \subseteq \mathcal{I}$ and $J \subseteq \mathcal{I} \setminus I$.

An itemset $I \subseteq \mathcal{I}$ is a *δ-free itemset* iff there is no δ-strong rule that holds between its proper subsets. In other words, $\forall S \subset I, |freq(S, r) - freq(I, r)| > \delta$.

When $\delta = 0$, δ is omitted, and we talk about *strong rules*, and *free itemsets*.

First introduced in [8], δ-free itemsets have been designed as an approximate condensed representation for frequency queries. δ-freeness is a generalization of the key pattern concept [4] (case $\delta = 0$). Let us discuss δ-freeness in terms of equivalence classes.

Definition 5 (*δ-closure, δ-closure equivalence class*) Let δ be an integer. The *δ-closure* of an itemset I on r is $cl_\delta : \mathcal{P}(\mathcal{I}) \rightarrow \mathcal{P}(\mathcal{I})$ s.t. $cl_\delta(I, r) = \{i \in \mathcal{I} \mid freq(I, r) - freq(I \cup \{i\}, r) \leq \delta\}$. When $\delta = 0$, $cl_0(I, r) = \{i \in \mathcal{I} \mid freq(I, r) = freq(I \cup \{i\}, r)\}$, and it corresponds to the well-known closure operator. We can also group itemsets by *δ-closure equivalence classes* (δ-CECs): two δ-free itemsets I and J are said δ-equivalent ($I \sim_{cl_\delta} J$) if $cl_\delta(I, r) = cl_\delta(J, r)$.

When $\delta = 0$, we get the formalization of closure equivalence classes from [4]. We can also derive δ-strong rules from δ-CECs (i.e., from δ-free itemsets and their δ-closures). We have then a δ-strong association rule between a δ-free itemset and each element of its δ-closure. Now, again in Fig. 1, interesting δ-CECs are still Cases 3 and 4 if we consider that X is a γ-frequent δ-free itemset and c_i an element of its δ-closure. Once again, for the sake of minimality and non-redundancy, we prefer δ-free itemsets to their supersets.

Example Let us consider the data from Table 1, a frequency threshold $\gamma = 3$, and an error threshold $\delta = 1$. For example, itemset A is γ-frequent δ-free. Its δ-closure is made of items C, D, and class value c_1. The rule $A \rightarrow c_i$ is a δ-strong rule concluding on a class attribute.

In [5,44], δ-free itemsets and their associated δ-closures are combined to define the so-called δ-bi-sets—which appear as an extension of formal concept for noisy data. δ-bi-sets are examples of maximal combinatorial rectangles of 1 values having at most δ zeros per column. δ-free based bi-sets have been successfully applied to characterize clusters in noisy environments. In the following, we argue that δ-CECs are suitable for classification tasks by discussing its fair properties.

5 Properties of δ-CECs

We show that mining δ-CECs brings more information for classification tasks than the other approaches discussed in Sect. 3. We also establish that, under some conditions on γ and δ values, δ-strong rules capture the needed discriminative power.

Table 2 Contingency table for an association rule $X \rightarrow c$ concluding on a class attribute c

$X \rightarrow c$	c	$\bar{c}$	Σ
X	f_{11}	f_{10}	f_{1*}
$\bar{X}$	f_{01}	f_{00}	f_{0*}
Σ	f_{*1}	f_{*0}	f_{**}

5.1 Information content in δ-CECs

Let us consider the contingency table (Table 2) for an association rule $\pi : X \rightarrow c$ concluding on a class attribute c. Class distribution is denoted by f_{*1} and f_{*0} and is known for all approaches, the same for f_{**}. However, if we consider the proposals from [17,27] based on frequent closed itemsets mined per class, we only get directly the value f_{11} (i.e., $freq(X \cup c, r)$) and the value for f_{01} can be inferred. Closure equivalence classes in [38] only inform us on f_{1*} (i.e., $freq(X, r)$) and f_{0*}. In our approach, when mining γ-frequent δ-free itemsets whose δ-closure contain a class attribute (i.e., corresponding to Cases 3 and 4), f_{1*} and f_{10} (resp. the effective frequency of X and the number of errors made by π) are known. Thus, f_{11} and frequencies in other cells can be inferred. The knowledge of f_{11} and f_{10} is precious since most of interestingness measures used in pattern-based classification techniques are based on these two values, i.e., the frequency of X in r_c and its frequency in the rest of the data. That is why, other approaches need a post-processing phase. In the following, we see that we can avoid it.

5.2 Fair properties of δ-CECs

According to the formalization from [20], $\pi : X \rightarrow c_i$ is a δ-strong Characterization Rule (δ-SCR) if c_i is a class attribute and body X is minimal. X is minimal if there is no other frequent[1] rule $\pi' : Y \rightarrow c_i$ s.t. $Y \subseteq X$ and $conf(\pi', r) \geq 1 - \frac{\delta}{\gamma}$.

This enables to avoid classification conflicts in the set of mined δ-SCRs under simple conditions on δ and γ values: if $\delta \in [0; \lfloor \gamma/2 \rfloor[$ then the set of δ-SCRs cannot contain two conflicting rules; i.e., if $\pi : X \rightarrow c_i$ has been mined, then $\pi' : Y \rightarrow c_j$ with $j \neq i$ and $Y \subseteq X$ cannot exist (this is called a body conflict in [20]).

However, defining δ-SCR only based on the confidence measure is not sufficient for prediction, as a δ-SCR $\pi : X \rightarrow c_i$ with high confidence does not say that X is positively correlated with c_i [10]. Therefore, we propose to exploit the growth rate measure Gr—an interestingness measure characterizing emerging patterns (EPs), i.e., patterns that are frequent in a class and barely infrequent in the rest of the database. Gr and EPs have been already proved useful for classification tasks (see e.g. [22]). The growth rate of $\pi : X \rightarrow c_i$ is defined as a ratio of relative frequencies as follows:

$$Gr(\pi, r_{c_i}) = \frac{freq_r(X, r_{c_i})}{freq_r(X, r \setminus r_{c_i})}$$

where r_{c_i} is the database restricted to objects of class c_i. Given a growth rate threshold $\rho > 1$, we talk about ρ-EPs. Using Gr to define δ-SCR ensures that X is positively correlated with c_i (see Proposition 1 and its proof in Appendix).

[1] Here, frequent means $(\gamma - \delta)$-frequent since γ denote the frequency of rule body.

Table 3 Contingency table for a δ-strong rule $X \to \delta\ c_i$ w.r.t. γ and δ values

$X \to c_i$	c_i	$\bar{c_i}$	Σ
X	$\gamma - \delta$	δ	γ
$\bar{X}$	$\cdot$	$\cdot$	$\cdot$
Σ	$\lvert r_{c_i}\rvert$	$\lvert r \setminus r_{c_i}\rvert$	$\lvert r\rvert$

Proposition 1 *Let ρ be a positive integer expressing a growth rate threshold such that $\rho > 1$ and let $\pi : X \to c_i$ an association rule concluding on a class attribute c_i. Then,*

$$Gr(\pi, rc_i) > \rho \quad \Longrightarrow \quad X \text{is positively correlated with } c_i$$

In [33], Gr, $conf$, and other interestingness measures are set in the general framework of the so-called δ-dependent measures. Such measures depend on the rule antecedent frequency (γ) and the rule number of exceptions (δ) following two principles:

(i) When γ is fixed, $Gr(\pi, r)$ increases with $freq(\pi, r)$;
(ii) When δ is fixed, $Gr(\pi, r)$ increases with γ.

This leads us to lower bounds for several interestingness measures (including growth rate and confidence) w.r.t. γ and δ values (see [33] for details). In Table 3, the contingency table for $\pi : X \to c_i$ which is a δ-strong rule concluding on class attribute c_i shows that, by construction, we have a lower bound $(\gamma - \delta)$ for $freq(X, r_{c_i})$, an upper bound δ for $freq(X, r \setminus r_{c_i})$, and other derivable bounds for dotted cells. Moreover, we can deduce a lower bound for the Gr and $conf$ measures. Indeed,

$$Gr(\pi, r) \geq \frac{\gamma - \delta}{\delta} \cdot \frac{|r \setminus r_{c_i}|}{|r_{c_i}|} \quad \text{and} \quad conf(\pi, r) \geq 1 - \delta/\gamma$$

Using such bounds, we can enhance the definition of δ-SCRs with the growth rate measure.

Definition 6 (*δ-strong characterization rule*) Let γ be a frequency threshold and $\delta > 0$ an error threshold. $\pi : X \to c_i$ is a *δ-strong characterization rule* (δ-SCR) if c_i is a class attribute, and the body X is minimal. We say that X is *minimal* if there is no other frequent rule $\pi' : Y \to c_i$ s.t. $Y \subseteq X \wedge conf(\pi', r) \geq 1 - \frac{\delta}{\gamma} \wedge Gr(\pi, r_{ci}) \geq \frac{\gamma-\delta}{\delta} \cdot \frac{|r \setminus r_{c_i}|}{|r_{c_i}|}$.

So, under some conditions on γ and δ values, and given a growth rate threshold ρ, this formalization ensures that bodies of δ-SCRs are ρ-EPs. These sufficient conditions are reported through the following proposition (see its proof in Appendix).

Proposition 2 *Let γ be a frequency threshold, $\delta > 0$ be an error threshold, and $\rho > 1$ be a growth rate threshold. Consider $\pi : X \to c_i$ a δ-strong rule s.t. X is a γ-frequent δ-free itemset and c_i a class attribute. Then,*

$$\delta < \frac{\gamma}{\rho} \cdot \frac{|r \setminus r_{c_j}|}{|r|} \quad \Longrightarrow \quad Gr(\pi, r_{c_i}) \geq \rho, \text{thus} X \textit{ is a } \rho\text{−EP} \tag{1}$$

$$\delta < \gamma/2 \quad \Longrightarrow \quad conf(\pi, r) \geq 1/2 \tag{2}$$

where c_j is the majority class.

Thus, for a given frequency threshold γ, γ-frequent δ-free itemsets X whose δ-closures contain a class attribute c_i such that δ satisfies Eqs. (1) and (2) from Proposition 2 are ρ-EPs without body conflicts. In the rest of the paper, γ and δ values are constrained w.r.t. Eqs. (1) and (2). We now consider how we use δ-SCRs to derive robust features.

6 A feature construction process

Our key idea is to build a new descriptor based on each body of extracted δ-SCRs. The new descriptor is simply a new attribute. In [17], a binary encoding is used. For a transaction t, the value of a new descriptor is 0 or 1 whether the corresponding δ-SCR match or not with t. Note also that recent advances on matrix decomposition meet this way of building features—see e.g. [42] where a feature is an itemset, and transactions may be re-mapped as a set of features. In this section, we propose and motivate a more promising numeric encoding to build new features. This feature construction process is summarized within Algorithm 1.

Algorithm 1: Building new data set with pattern-based new descriptors

```
input  : A binary database $r = \{\mathcal{T}, \mathcal{I}, \mathcal{R}\}$, two integers $\gamma$ and $\delta$ as thresholds for frequency and errors
output: A new numeric database $r' = \{\mathcal{T}, \mathcal{I}', \mathcal{R}'\}$ made of pattern-based descriptors

1 begin
2     $\mathcal{I}' \longleftarrow$ FeaturesExtraction($r, \gamma, \delta$);
3     for $t \in \mathcal{T}$ do
4         for $I' \in \mathcal{I}'$ do
5             $\mathcal{R}'(t, I') \longleftarrow \frac{|I' \cap \mathrm{Items}(t,r)|}{|I'|}$;
6     $r' \longleftarrow \{\mathcal{T}, \mathcal{I}', \mathcal{R}'\}$;
7 end
```

Extraction step. Procedure `FeaturesExtraction` (Line 2) mines all γ-frequent δ-free itemsets I' that are bodies of δ-strong characterization rules in r. This step is performed efficiently using a straightforward extension the level-wise algorithm presented in [8]. It benefits from the anti-monotonic properties of δ-freeness and γ-frequency constraints to compute all the γ-frequent δ-free itemsets. Moreover, since we are interested in minimal itemsets whose δ-closures contain a class attribute, the two following constraints have been added:

- $\mathbb{C}_1 \equiv \exists c$ (class attribute) $| \ c \in cl_\delta(X, r)$ (syntactic constraint)
- $\mathbb{C}_2 \equiv \nexists Y \in S_{\gamma,\delta} \mid Y \subseteq X$ (minimal body constraint)

Additional constraints $\mathbb{C}_1$ and $\mathbb{C}_2$ contribute to the extraction of a smaller feature set. Indeed, only class-discriminant features are extracted ($\mathbb{C}_1$), and there is no need to check supersets of itemsets selected in a precedent level ($\mathbb{C}_2$).

Construction step. Then, each I' becomes a new descriptor for r', and (Line 5) the value of I' for a transaction t is the proportion of items in I' that are supported by t in r. As we said earlier, $Items$ is the dual operator for $Objects$. Now, $\mathcal{R}' \mapsto [0;1]$ and $\mathcal{R}'(t, I') \in \left\{0, \frac{1}{|I'|}, \ldots, \frac{|I'|-1}{|I'|}, 1\right\}$. We think that multi-valued encoding—followed by, e.g., an entropy-based splitting supervised discretization step—should preserve more information than binary encoding. Indeed, in the worst case, the split will take place between $\frac{|I'|-1}{|I'|}$ and 1, that is equivalent to the binary case. Otherwise, the split may take place between $\frac{j-1}{|I'|}$ and $\frac{j}{|I'|}$, $1 \leq j \leq |I'| - 1$, and this split leads to a better separation of the data. Finally, r' is the new database made of noise-tolerant features, ready for a classifier learning step. The generic processus schema in Fig. 2 reminds the two main steps—constraint-based mining and feature construction. Several instances of such a processus are empirically studied in the next section.

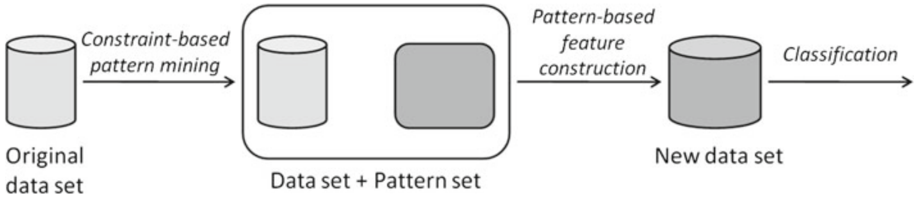

Fig. 2 Generic processus for feature construction

Complexity of FC process. The extraction step is obviously the hardest one. The original frequent itemset mining task is computationally hard since there may be a huge (exponential) number of frequent itemsets. In [9], the authors show that the number of frequent δ-free itemsets is significantly smaller than for the frequent ones. During the FC process, when mining δ-SCRs, constraints $\mathbb{C}_1$ and $\mathbb{C}_2$ significantly reduce the search space and the size of the output feature set. Thus, the extraction of δ-SCRs is tractable with the proposed level-wise algorithm in most of the labeled data sets that are not too large w.r.t. the number of attributes. Notice that, in the special case of "fat" data sets such as gene expression data sets, dedicated approaches have been suggested (see, e.g., [32]).

7 Empirical study

We performed several experimentations to validate our feature construction process (FC). These experimentations are used to answer the following questions:

- Q_1 Given a *classical*[2] classification technique (e.g., C4.5 decision tree induction [45], Naïve Bayes rule NB [34], support vector machine SVM), can we tell that the obtained classifier is more accurate when using the new features (generated thanks to FC) than when using the original attributes?
- Q_2 What happens if the original attributes are noisy? Does FC enable to predict better than when using the original attributes? Do we gain some noise-tolerance? How can we set the δ values when the attributes are noisy?
- Q_3 When using the FC processus followed with a classical technique, how do we compare with an efficient pattern-based method like HARMONY [53,54]?

7.1 Experimentation protocol

In order to answer those questions, we set up two protocols that deal with the original data and artificially noisy data.

Original data: We performed our feature construction process (denoted FC) on several UCI data sets (see a brief description in Table 4). We used C4.5 decision tree induction, Naive Bayes classification (NB), and SVM on both the original data sets and their transformations thanks to Algorithm 1.

Attribute-noisy data: We also experimented our process in attribute-noisy data. In this context, we want to learn accurate predictive models despite of noisy samples. Therefore, in our experiments, we deal with attribute-noisy training sets and clean (say noise-free) test sets. We added random noise at different levels only on attributes and only in training sets. For

[2] We identify as classical three classification algorithms from the top 10 algorithms in data mining [56].

Table 4 Used UCI benchmark data sets

Dataset	#Instances	#Attributes	#Classes
breast-w	699	9	2
colic	368	22	2
diabetes	768	8	2
heart-c	303	13	2
heart-h	294	13	2
heart-s	270	13	2
hepatitis	155	19	2
iris	150	4	3
tic-tac-toe	958	9	2
wine	178	13	3
vote	435	17	2

a data set d and a $x\%$ noise level ($x \in \{10, 20, 30, 40, 50\}$), each attribute value got a $x\%$ probability to be changed (within its range values including its current value) in each transaction of the training set.[3] When dealing with continuous attributes, we first discretized *only* the training data. Each attribute was split into several intervals using entropy-based splitting method [24]. Then, we added $x\%$ noise and we performed a simple binarization. Discretization and binarization are finally reported on test data. All pre-processing steps (adding noise, discretization, binarization) and accuracy results for NB, C4.5 and SVM classifiers are obtained with 10-fold stratified cross-validation—using the Weka platform [55] with LibSVM and WLSVM libraries [16,23].

7.2 A strategy for setting δ

Our FC process depends on two parameters for the frequency γ and the number of errors δ. We know that using extreme γ values is not relevant. Lowest values lead to a huge amount of rules, a lot of them bring poor information since they are supported by few samples. Highest values lead to few rules and such that the training data may not be reasonably covered. Automatically setting frequency threshold is still an open question (see [15,59] for preliminary results in that direction).

Given a frequency threshold γ, how can we determine relevant δ values? Evolution of δ-dependent interestingness measure (such as growth rate or confidence) w.r.t. δ is well known. Decreasing δ implies higher values for Gr, but such interesting patterns could be rare, especially in noisy data sets. When increasing δ, extracted patterns tend to match with noisy patterns in training data, but higher δ values tend to be less relevant (with low Gr values) since a lot of errors are allowed. We propose a simple strategy for setting δ values. We motivate this strategy with an experimental study of the evolution of accuracy results w.r.t. δ values.

In Figs. 3, 4 and 5, we plotted accuracy results w.r.t. δ values for different frequency thresholds and noise levels, for FC-C4.5 on tic-tac-toe data set (Fig. 3), for FC-NB on colic (Fig. 4), and for FC-SVM on heart-c data set (Fig. 5). First, we remark that the accuracy increases (not necessarily monotonically) with δ until a maximal point—often better

[3] It does not mean that $x\%$ of values have been modified.

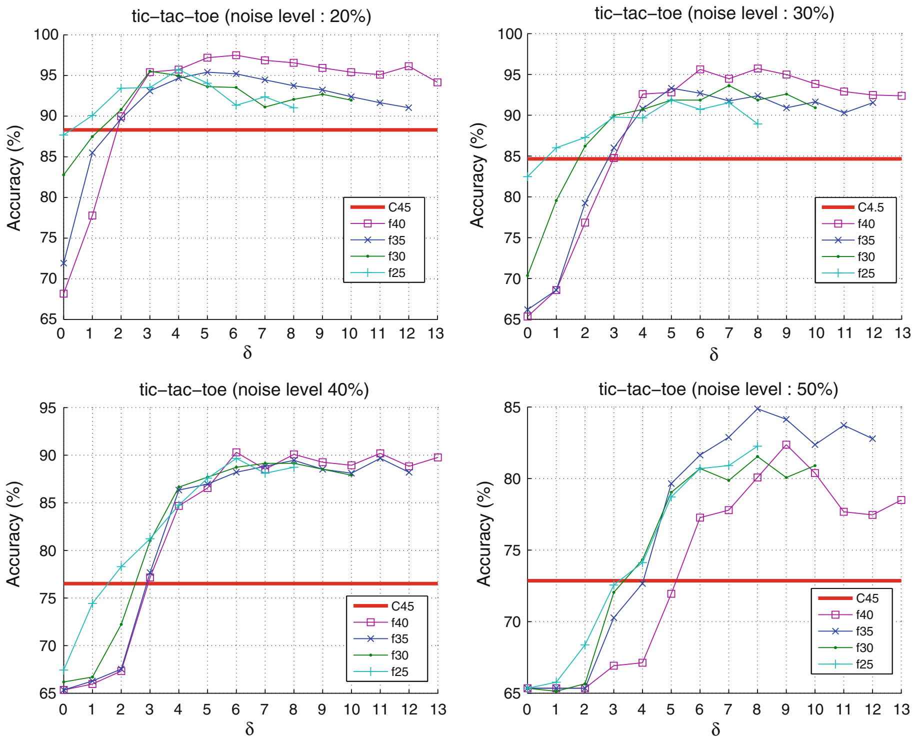

Fig. 3 Evolution of accuracy for FC-C4.5 w.r.t. δ values for different noise levels and frequency thresholds on tic-tac-toe data set

than accuracy obtained with noisy data sets and original attributes (see thick red line). Then, accuracy decreases. This is the general behavior of accuracy w.r.t. δ, and it is emphasized when noise increases. But in some cases, accuracy decreases faster for NB. Indeed, C4.5 sequentially chooses the best itemset for splitting and thus building the decision tree. Not necessarily all extracted itemsets are used in the tree. In the contrary, all extracted itemsets are used in the product approximations within a NB classifier.

In Fig. 6, we plotted accuracy results w.r.t. noise level for various δ values (one per curve) and various γ values (one per graph). We remark that, for low levels of noise, FC with low δ values are sufficient to ensure good accuracy results compared to the original classifier (here C4.5). As the noise level increases, FC with low δ values—which produces *almost* strong rules—gives poor accuracy results. Obviously, trying to find strong correlations in noisy data is inefficient. Conversely, FC with higher δ values seems to match better with noisy patterns than with low values. Note also that FC can achieve better accuracy results for several γ thresholds.

Since we have a better understanding of relations between γ, δ, noise levels, and their impact on classification results, let us propose a way for setting δ values. In Fig. 7, we plotted training accuracy results w.r.t. δ values for tic-tac-toe data set. As expected, accuracy on training data increases with δ until stabilization (or slowing down, or decreasing, depending on the data set). δ values around the stabilization area are interesting since lower values lead to less accurate models and higher values bring nothing more. Let δ_{opt} denote these values. These interesting values depend on the amount of noise in data. In Fig. 7, we see that

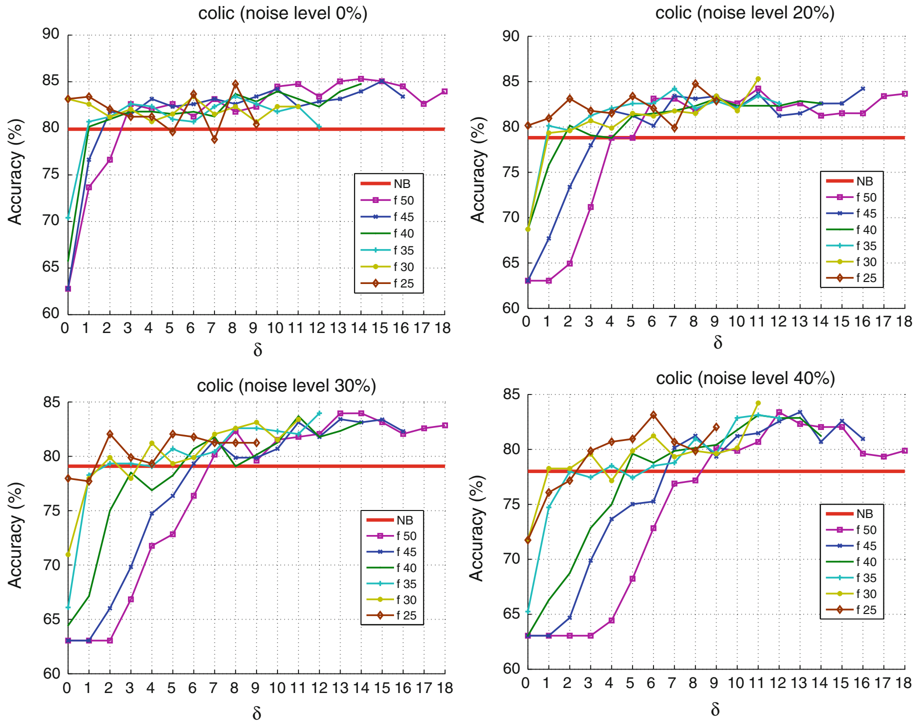

Fig. 4 Evolution of accuracy for FC-NB w.r.t. δ values for different noise levels and frequency thresholds on colic data set

as noise level increases, δ_{opt} value increases too. Since, in real case, noise level is not known a priori, a reasonable way to reach these δ_{opt} values is:

(i) Increase δ starting from 0;
(ii) Check evolution of training data accuracy for stabilization (or decreasing).

Note that in our previous work [29,30], we proposed to check for δ_{opt} values using the evolution of training data coverage—not the training accuracy. We think that following training accuracy evolution is more relevant. Indeed, if you consider a set S of new descriptors generated by FC , then training data coverage proportion is the same whatever the classifier ending FC process, and so are our δ_{opt} values. Unfortunately, we experimentally remarked that *good* δ values are not necessarily the same when using different classification methods after FC . This is due to the way a classifier works. That is why, we prefer to look after training accuracy results which enables different δ_{opt} for the used classifier.

7.3 Accuracy results

Accuracy result comparison is shown in Table 5. We want to make two comparisons. First, we compare accuracy results of each classifier (computed thanks to C4.5, NB, or SVM with radial basis function as kernel) on original data versus itself on data enhanced by FC on original and noisy data. Next, we compare the result of the FC process w.r.t. HARMONY. For each data set, we used different γ-frequency thresholds. For each γ, we applied our strategy to set δ. FC accuracy results are reported in two columns: *Avg* for average accuracy overall

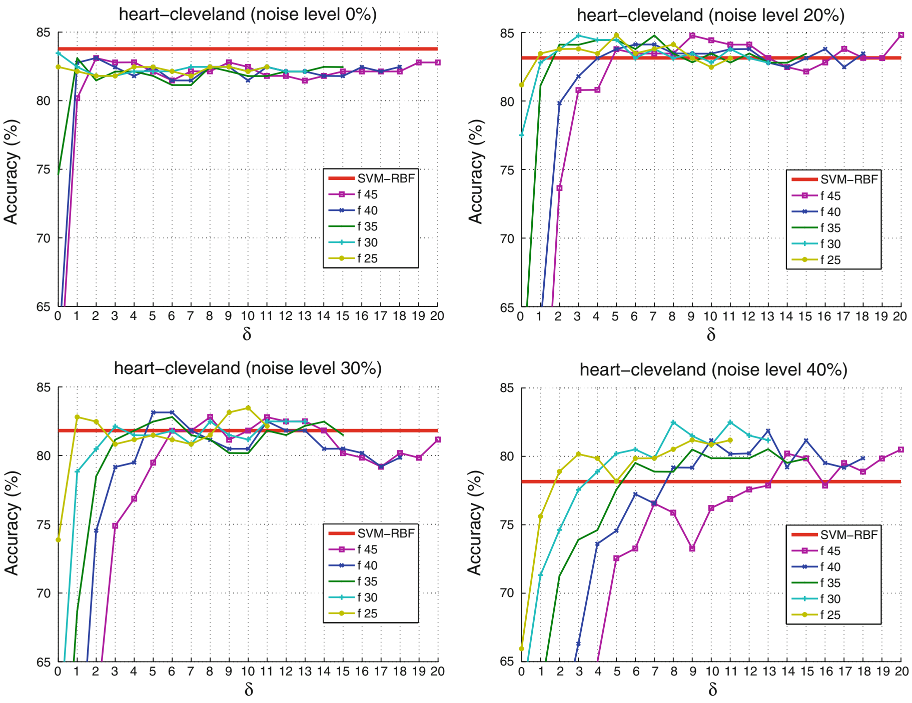

Fig. 5 Evolution of accuracy for `FC-SVM` w.r.t. δ values for different noise levels and frequency thresholds on another data set

tested γ values for δ_{opt} values and Max for the maximal accuracy over all γ values and corresponding δ_{opt}. Note that the same cross-validation on the same noisy data is used for accuracy comparison.

First, we remark that when the data are not artificially noisy, `FC-C4.5` and `FC-NB` often achieve better accuracies than the original classifiers (10/11 for `C4.5` and 7/11 for `NB`). If we consider a *good* γ threshold (column Max), these ratios become respectively 11/11 and 8/11. Results for `SVM-RBF` are not so good (4/11 and 7/11 with a selected γ). However, average accuracy improvement of `FC-SVM` is insignificant. We think the non-tuning of manifold parameters of `SVM-RBF` technique could be one of the reasons of `FC-SVM`'s poor results.

When the data are noisy, `FC` often works well and classifiers learnt on enhanced data sets are more accurate: `FC-C4.5` wins 35 times over 55 versus `C4.5`. For `FC-NB` and `FC-SVM`, the ratios are respectively 28/55 and 40/55 (bold faced results). Once again, if we consider a relevant γ threshold in particular, then we have better ratios (50/55, 41/55, and 42/55). At the end of Table 6, we report average improvement over all data sets tested. We clearly see that, on noisy data sets, it is worth using `FC` to learn more accurate classifiers in most of the cases. The exception is perhaps for `NB`. Results showed that `NB` is *naturally* noise-tolerant, so `FC-NB` improvement are less impressive and sometimes (for high level of noise), `FC-NB` deteriorates performance, especially if γ is not well selected. Average accuracy improvement results show that `FC` results are more stable when `C4.5` ends the process. Decision tree induction intrinsic sequential selection of features is probably at work here. An interesting extension—not developed in this paper—could be a post-selection of our features before piping `NB` and `SVM`.

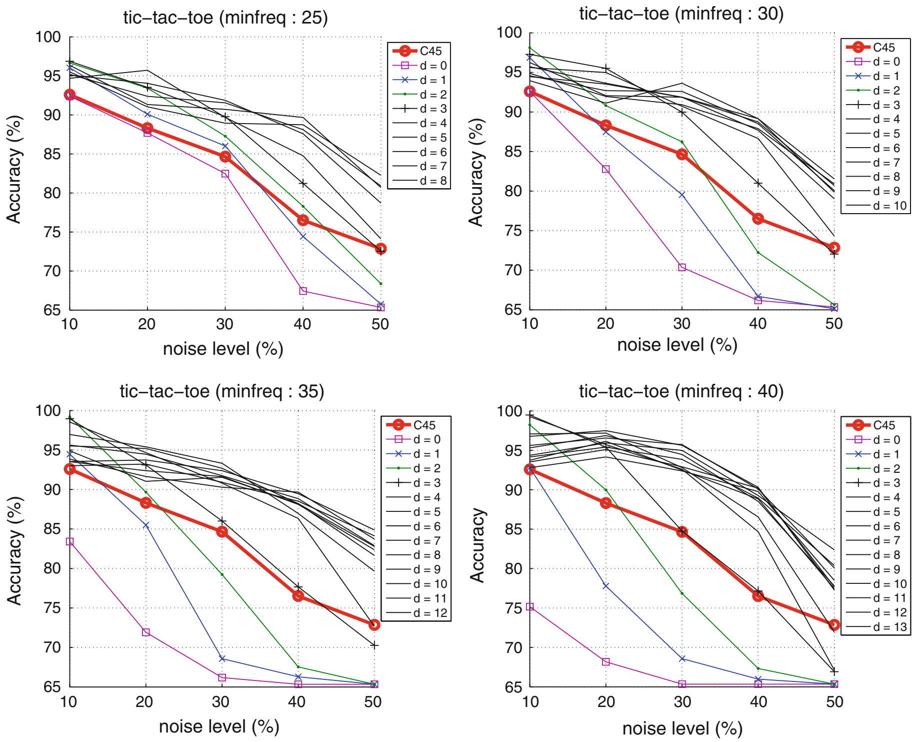

Fig. 6 Evolution of accuracy w.r.t. noise level for different frequency and number of errors thresholds on `tic-tac-toe` data set

As the authors from [54], we produce accuracy results for HARMONY with different frequency thresholds (5,10,15), and we report the best. We then compare HARMONY with the Max column of FC processes. Original C4.5 (resp. NB and SVM) wins 9 times (resp. 37 and 33) over 66 versus HARMONY. Except for NB that benefits from its natural noise-tolerance, HARMONY is clearly better than C4.5, and it is comparable with SVM. When processing FC, we noted how many times an original classifier loses versus HARMONY then wins using FC : FC-C4.5 scored 23, FC-NB 12 and FC-SVM 9. Once again, it is worth using FC. A classifier computed by FC-C4.5 becomes comparable with the classifiers delivered by HARMONY (see Table 7 for average results). Indeed, FC-C4.5 wins 32 times over 66; FC-NB 44/66 and SVM 40/66.

8 Discussion and related work

Limits. A clear limit encountered by our approach is when class distribution is unbalanced. In fact, our FC approach is based on a global frequency threshold γ. To get a chance to characterize a minor class, low γ value may be chosen and since δ is constrained w.r.t. γ, the number of errors allowed δ may also be very low (perhaps 0). Therefore, strong rules induced by such γ and δ values may be inefficient for classification purposes. Note that all CBA-like techniques (i.e., classification based on association rules) using a global frequency threshold suffer from this problem. Conversely, HARMONY follows an instance-centric principle

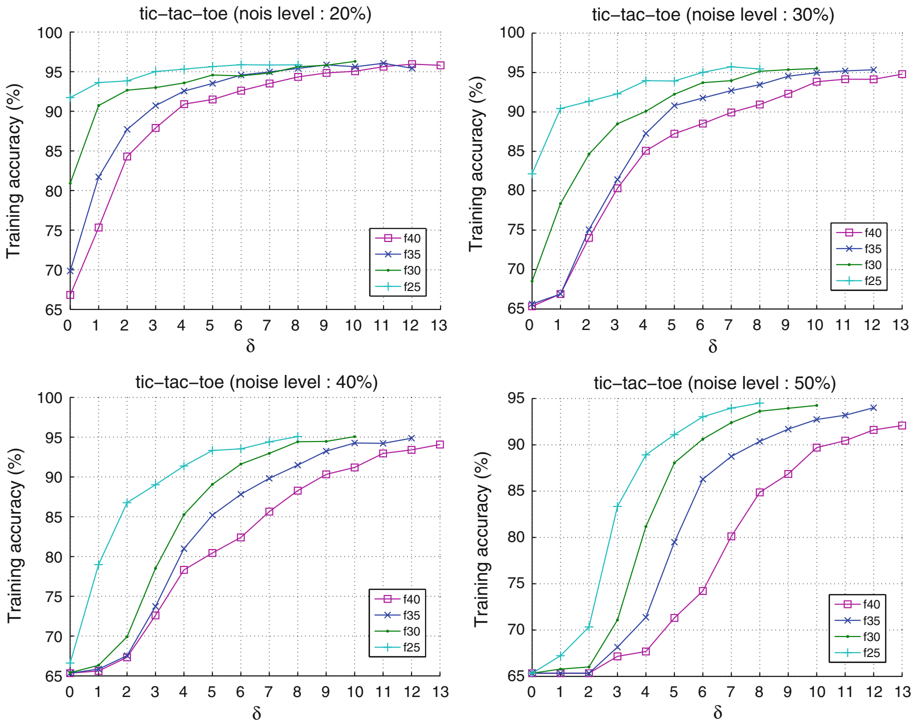

Fig. 7 Evolution of `FC & C4.5` accuracy on training database w.r.t. δ for various frequency thresholds and noise levels on `tic-tac-toe` data set

and thus optionally allows a specific frequency threshold per class. Specific thresholds are set using *support differential factor* which is another parameter to be tuned in addition to minimum support of minor class.

Another limit of our approach (and all aforementioned approaches) is when facing multi-class problem. Using confidence or growth rate as driving interestingness measures lead us to `One-Versus-All` framework (OVA), where we try to characterize a class w.r.t. the rest of the data. In such a framework, the repartition of errors made by extracted interesting patterns in the various classes of the rest of the data is not taken into account. If our `FC` approach is forced to lower γ threshold to avoid conflicts, in other `CBA`-like or EPs-based approaches, it could happen that some extracted patterns are positively correlated with more than one class—resulting in classification conflicts. There are two other frameworks dedicated to multi-class problems: (i) the `One-Versus-One` framework (OVO), in which a multi-class problem (say $n > 2$ classes) is divided into $n(n-1)$ 2-class subproblems. Then, classifiers learnt for the various subproblems are combined thanks to, e.g., a voting scheme. This framework has given rise to several approaches known as round robin classification [25] or pairwise classification [43]. (ii) In [15], the authors present another approach to deal with unbalanced multi-class distribution using various local thresholds: given a n-class problem, for each class c_i, they extract patterns w.r.t. a specific support threshold for c_i and a specific error threshold for each $c_j \neq c_i$. It can be seen as an `One-Versus-Each` approach (OVE). The authors propose a hill-climbing strategy to automatically set the n^2 parameters. It results in a parameter-free associative classifier dedicated to unbalanced distribution

Table 5 Accuracy results: original versus FC processed data

Data sets	C4.5	FC-C4.5	Max	NB	FC-NB	Max	SVM	FC-SVM	Max	HARMONY
breast-w (%)	95.57	**95.62**	**95.85**	97.28	96.59	96.85	96.85	**97.19**	**97.42**	95.85
10	95.14	94.59	**95.14**	96.99	**97.22**	**97.56**	96.56	**97.19**	**97.42**	95.71
20	93.28	92.31	92.99	96.99	**97.05**	**97.14**	95.85	*95.82*	**96.14**	95.71
30	89.56	**91.00**	**91.85**	96.85	**97.14**	**97.28**	95.14	**95.16**	**95.42**	94.99
40	87.70	**87.73**	**88.41**	96.42	94.99	**97.28**	92.56	91.28	91.85	93.56
50	87.56	86.30	**87.56**	96.28	**96.51**	**97.21**	86.41	83.06	83.98	90.99
colic (%)	85.04	84.31	**85.85**	79.90	**82.73**	**85.31**	84.77	84.59	**84.77**	82.88
10	83.14	82.95	**85.29**	78.81	**82.59**	**84.77**	84.50	**85.17**	**86.12**	82.06
20	82.04	80.88	**83.12**	78.81	**81.28**	**85.31**	83.94	**84.88**	**85.82**	81.79
30	80.95	80.78	**82.85**	79.09	**81.37**	**83.96**	80.98	**84.61**	**85.28**	82.88
40	80.93	**81.72**	**84.21**	78.00	**82.32**	**84.22**	69.55	**81.87**	**85.30**	83.15
50	81.53	**82.04**	**83.66**	73.92	**80.56**	**82.31**	63.05	**76.14**	**81.51**	82.61
diabetes (%)	73.58	**73.96**	**74.23**	73.96	*73.61*	**75.00**	74.74	**75.05**	**75.56**	73.05
10	72.79	**72.79**	**73.31**	73.31	72.69	**73.83**	73.44	**75.19**	**75.39**	73.70
20	72.02	**72.02**	**72.27**	73.83	72.50	73.32	72.92	71.58	72.61	73.70
30	71.23	**72.02**	**73.06**	73.44	72.56	**73.70**	72.01	65.24	65.62	72.66
40	68.88	**69.57**	**70.96**	75.13	71.76	72.97	70.06	65.11	65.11	70.96
50	68.37	67.94	**69.15**	70.45	67.55	68.75	67.59	65.11	65.11	68.10
heart-c (%)	80.47	**82.48**	**84.13**	81.79	**83.20**	**84.12**	83.77	**83.88**	**84.10**	82.18
10	78.83	**80.17**	**81.18**	82.78	**83.39**	**84.78**	83.12	**84.64**	**85.09**	81.19
20	75.85	**79.46**	**82.18**	83.12	**83.26**	**84.46**	82.48	**84.59**	**84.79**	80.86
30	77.19	**78.40**	**80.21**	84.45	83.93	**84.82**	81.82	**83.13**	**83.79**	82.84
40	77.50	76.40	**78.52**	84.45	82.63	**84.48**	79.16	**81.90**	**82.84**	78.88
50	72.28	**75.53**	**77.52**	79.52	**79.93**	**88.20**	77.48	**79.14**	**80.51**	76.24
heart-h (%)	75.55	**79.60**	**81.64**	84.07	82.50	83.03	81.38	**82.49**	**83.38**	82.31
10	77.62	**78.83**	**80.56**	83.73	80.70	81.31	82.02	**83.31**	**83.71**	84.01
20	76.60	**78.24**	**80.29**	83.39	82.30	83.00	82.70	81.88	82.67	80.61
30	75.56	**79.08**	**80.00**	84.05	83.10	**84.34**	81.32	**82.34**	**82.67**	79.93
40	74.17	**79.19**	**80.96**	83.37	80.51	82.00	80.29	**82.40**	**83.01**	81.63
50	73.83	**79.93**	**81.68**	79.95	**80.17**	**82.34**	68.63	**81.83**	**82.00**	78.57
heart-s (%)	81.85	**83.33**	**86.67**	81.48	**82.10**	**84.45**	83.33	83.18	**83.70**	81.48
10	80.74	**81.17**	**84.45**	81.48	81.17	**85.92**	83.70	83.40	**84.07**	81.85
20	78.52	**79.75**	**80.74**	82.59	81.60	**84.07**	80.00	**83.27**	**84.45**	80.74
30	74.07	**74.07**	**76.30**	81.11	**81.92**	**84.07**	78.15	**83.52**	**84.81**	77.41
40	72.96	70.15	**73.33**	81.85	71.70	*72.59*	72.59	58.81	63.70	70.74
50	59.26	**64.07**	**65.55**	55.55	**61.18**	**65.18**	59.63	55.55	55.55	63.70
hepatitis (%)	81.87	**82.71**	**85.17**	82.00	**82.22**	**84.50**	83.21	79.69	81.33	83.87
10	80.62	80.19	**81.33**	83.25	82.90	**84.50**	80.67	79.37	79.37	83.87
20	78.62	**78.93**	**79.83**	85.21	84.46	**86.42**	79.37	79.37	79.37	81.29
30	79.29	78.23	**81.08**	86.46	83.81	84.37	79.37	**79.37**	**79.37**	81.93
40	76.04	**78.41**	**79.79**	85.79	82.34	83.79	79.37	**79.37**	**79.37**	78.06
50	80.58	79.46	**83.79**	79.37	**79.81**	**81.21**	79.37	**79.37**	**79.37**	81.93

Table 5 continued

Data sets	C4.5	FC-C4.5	Max	NB	FC-NB	Max	SVM	FC-SVM	Max	HARMONY
iris (%)	93.33	**93.33**	**93.33**	92.67	**94.00**	**94.00**	94.67	94.00	94.67	95.33
10	92.67	**93.00**	**94.67**	92.67	**92.67**	**94.00**	94.67	94.00	**94.67**	90.00
20	90.00	89.22	**93.33**	92.67	**93.11**	**94.67**	94.00	**94.00**	**94.67**	90.67
30	90.00	86.50	89.33	92.67	91.33	92.00	92.00	**92.00**	**94.67**	91.33
40	86.00	84.67	84.67	92.67	88.00	88.33	90.67	88.80	**90.67**	90.00
50	84.00	77.33	77.33	94.00	77.33	77.33	90.67	**80.77**	**84.00**	82.00

Table 6 Accuracy results: original versus FC processed data (next)

Data sets	C4.5	FC-C4.5	Max	NB	FC-NB	Max	SVM	FC-SVM	Max	HARMONY
t-t-t (%)	93.21	**99.40**	**100.00**	68.47	**79.72**	**81.73**	87.79	86.40	89.24	97.08
10	92.06	**97.65**	**99.17**	68.57	**75.99**	**78.81**	81.10	76.90	77.97	96.97
20	88.95	**96.26**	**97.18**	70.14	**73.81**	**74.42**	71.92	65.34	65.45	94.68
30	82.78	**93.00**	**95.62**	71.92	**73.56**	**74.53**	65.65	65.34	65.34	91.13
40	77.98	**89.43**	**91.55**	74.11	73.56	**74.83**	65.34	**65.34**	**65.34**	89.67
50	71.39	**83.82**	**85.60**	70.57	**71.16**	**71.39**	65.34	**65.34**	**65.34**	81.84
wine (%)	91.08	**91.53**	**93.76**	98.89	92.92	93.86	98.33	96.91	97.22	96.63
10	91.63	**91.87**	**94.38**	96.6	93.24	94.41	98.30	96.48	97.19	95.50
20	91.01	**91.74**	**92.71**	96.04	94.42	95.00	97.19	**97.19**	**97.19**	96.07
30	84.37	**88.15**	**91.04**	94.38	91.73	**94.97**	97.19	**97.19**	**97.74**	92.70
40	82.02	**88.28**	**88.82**	93.27	91.42	**93.79**	96.08	95.24	**96.63**	91.01
50	82.12	**85.11**	**87.58**	88.10	**89.29**	**90.42**	91.60	68.76	69.74	88.76
vote (%)	95.19	**96.32**	**96.78**	90.37	**92.14**	**92.20**	95.41	95.29	**95.41**	94.71
10	94.26	**94.53**	**94.94**	89.68	**92.20**	**92.44**	95.64	95.13	95.42	95.63
20	92.42	**93.78**	**94.24**	89.91	**91.18**	**91.59**	95.41	94.84	95.18	94.48
30	93.11	91.25	91.96	88.98	**90.77**	**91.28**	94.72	94.38	94.49	94.94
40	90.59	90.56	**91.72**	89.21	**90.71**	**91.74**	93.58	92.91	93.11	92.64
50	89.44	88.58	**89.21**	89.21	**90.19**	**90.36**	92.20	89.78	90.13	93.10
Average (%)		+1.44	+2.79		+0.99	+2.2		−0.51	+0.23	
10		+0.75	+2.27		+0.63	+2.22		−0.27	+0.25	
20		+1.21	+2.69		+0.21	+1.52		−0.27	+0.23	
30		+1.31	+3.20		−0.20	+1.08		+0.36	+0.99	
40		+1.94	+3.47		−2.21	−0.75		−0.57	+0.70	
50		+1.80	+3.48		−0.29	+1.62		−1.56	−0.43	

problems. However, due to the multiple extraction induced by the hill-climbing strategy, the computation time increases significantly.

Error-tolerant pattern mining. Since the proposal of the error-tolerant itemset mining task (ETI) in [57], many algorithms have been designed. If we focus on condensed representations of ETI based on almost-closedness concept, [5] has proposed several extensions of formal concepts to meet fault-tolerance. Our approach is based on one of this extension, namely the

Table 7 Average accuracy results of HARMONY and comparison with FC

Average (%)	HARMONY	C4.5	FC-C4.5 (max)	NB	FC-NB (max)	SVM	FC-SVM (max)
0	87.72	−1.65	+1.14	−3.09	−0.89	−0.06	+0.17
10	87.32	−1.91	+0.36	−2.97	−0.77	−0.62	−0.37
20	86.42	−2.85	−0.16	−1.63	−1.11	−1.35	−1.12
30	85.70	−4.05	−0.85	−0.85	+0.23	−2.21	−1.22
40	83.66	−4.14	−0.67	+1.27	+0.52	−2.82	−2.12
50	80.71	+0.08	−2.60	−0.99	+0.63	−4.17	−4.60

Free set based Bi-Set (FBS). We may point out two weaknesses of FBS in pattern mining tasks. First, the appearance of errors is limited to the attributes from the δ-closure. Next, an absolute error threshold is used. Among others, we may cite two other approaches like [18] and [19]. The one tolerates a relative frequency gap for an itemset to be closed. The other allows a relative number of errors (i.e., 0s) per line and per column in the Boolean matrix. Relative thresholds and no limitations on the columns where 0 values appear seem to be more suitable. Since both approaches focus on closedness and thus exploit the inherent maximality property of closedness, an interesting direction for work could be the study of an extension of [17]'s approach for noise-tolerant classification. Note that, here, more investigations are needed if one wants to build relevant features from these almost-closed itemsets since the derivation of class-characterization rules is not as natural as from a δ-CEC.

Noise-tolerant classification. When dealing with classification tasks, [62] has shown that the presence of noise in the data can have a negative impact on learnt classifiers. One may differentiate two main types of noise: class noise when noise affects the class label, and attribute noise when noise affects all attributes but not the class label. Many solutions have been proposed to tackle class noise, e.g., by noise elimination or noise correction (see [62] for a survey) and more recently by instance weighting [47]. Other approaches aim at solving the problem of attribute noise by noise identification and modeling [35,52,60], or noise cleansing [58,62]. In our FC approach, instead of removing noisy instances or correcting noisy values, we propose a method to cope with attribute noise without changing or removing any attributes values in the training data.

9 Conclusion

We studied the use of closedness-related condensed representations for feature construction. We pointed out that differences about "freeness or closedness" within existing approaches come from the way the patterns are mined, i.e., with or without class labels, per class or in the whole database. We then proposed an application-independent framework to construct features. Our new features are built from mined (δ)-closure equivalence classes – more precisely from γ-frequent δ-free itemsets whose δ-closures involve a class attribute. Mining these types of itemsets differs from other approaches. Among other things, we discussed the information content of such equivalence classes when γ and δ are carefully set. We also proposed a new numeric encoding that is more suitable than binary encoding for classification tasks. Our FC process has been validated by means of an in-depth empirical study. Using C4.5, SVM, and NB on new representations of various data sets, we demonstrated the accuracy improvement when using the new features instead of the original ones. In the particular case of attribute-noisy data, we highlighted that δ values are very important to introduce noise-tolerance in the feature construction process. We also showed comparable accuracy results w.r.t. the recent and efficient classification technique HARMONY. Finally, we now have a better understanding of the effects of almost-closedness properties on feature construction for classification purpose.

Acknowledgments We thank Jianyong Wang and George Karypis who provided the HARMONY prototype. This work has been partly funded by French contract ANR MDCO BINGO 2.

Appendix

In this section, we report proofs for Propositions 1 and 2 such that the article is self-contained.

Proof of proposition 1

To simplify notations, consider the contingency table for an association rule $\pi : X \rightarrow c_i$ (see Table 8). According to the *interest factor* [50], X is said to be positively correlated with c_i if

$$\widehat{IntF}(\pi, c_i) = \frac{a \cdot |r|}{(a+b)\cdot(a+c)} > 1$$

$$\text{i.e., } \frac{|r|}{a+b+c+\frac{b\cdot c}{a}} > 1$$

From another hand, given $\rho > 1$, X is a ρ-EP if

$$Gr(\pi, r_{c_i}) = \frac{a\cdot(b+d)}{b\cdot(a+c)} \geq \rho \text{ i.e., } a\cdot b + a\cdot d \geq \rho\cdot(a\cdot b + b\cdot c)$$

$$ab + ad \geq \rho ab + \rho bc \text{ then } b + d \geq \rho b + \rho\frac{bc}{a}$$

$$d \geq (\rho - 1)b + \rho\frac{bc}{a} > \frac{bc}{a} \text{ because } \rho > 1$$

$$\text{since } d > \frac{bc}{a} \text{ then } \widehat{IntF}(\pi, c_i) = \frac{|r|}{a+b+c+\frac{b\cdot c}{a}} > \frac{|r|}{a+b+c+d} = 1$$

And X is thus positively correlated with c_i. □

Proof of proposition 2

Given three thresholds γ, δ, and $\rho > 1$ (resp. for frequency, errors and Growth rate), we saw that to ensure the body X of a δ-SCR $\pi : X \rightarrow c_i$ to be a ρ-EP, it is enough that

Table 8 Contingency table for an association rule $X \rightarrow c_i$ concluding on a class attribute c_i

$X \rightarrow c$	c	$\bar{c}$	Σ
X	a	b	$a+b$
$\bar{X}$	c	d	$c+d$
Σ	$a+c$	$b+d$	$\|r\| = a+b+c+d$

$\frac{\gamma-\delta}{\delta} \cdot \frac{|r \setminus r_{c_i}|}{|r_{c_i}|} \geq \rho$. For that, it is enough that

$$\frac{\gamma-\delta}{\delta} \geq \frac{|r_{c_i}|}{|r \setminus r_{c_i}|} \cdot \gamma \text{ then } \frac{\gamma}{\delta} \geq \frac{|r_{c_i}|}{|r \setminus r_{c_i}|} \cdot \rho + 1$$

$$\text{i.e., } \frac{\gamma}{\delta} \geq \frac{\rho \cdot |r_{c_i}| + |r \setminus r_{c_i}|}{r \setminus r_{c_i}} \text{ then } \gamma \cdot \frac{|r \setminus r_{c_i}|}{\rho \cdot |r_{c_i}| + |r \setminus r_{c_i}|} \geq \delta$$

$$\text{Observe that } \frac{|r \setminus r_{c_i}|}{\rho \cdot |r_{c_i}| + |r \setminus r_{c_i}|} > \frac{\gamma}{\rho} \cdot \frac{|r \setminus r_{c_i}|}{|r_{c_i}| + |r \setminus r_{c_i}|}, \text{ then } \frac{\gamma}{\rho} \cdot \frac{|r \setminus r_{c_i}|}{|r_{c_i}| + |r \setminus r_{c_i}|} > \delta$$

Considering that there could be an unequal class distribution, it suffices that the precedent inequality is verified for the majority class. Then, $\frac{\gamma}{\rho} \cdot \frac{|r \setminus r_{c_j}|}{|r|} > \delta$ (c_j is the majority class) is a sufficient condition to ensure that X is a ρ-EP. □

References

1. Agrawal R, Imielinski T, Swami AN (1993) Mining association rules between sets of items in large databases. In: Proceedings ACM SIGMOD'93, pp 207–216
2. Antonie M-L, Zaïane OR (2004) An associative classifier based on positive and negative rules. In: Proceedings of the 9th ACM SIGMOD workshop on research issues in data mining and knowledge discovery, DMKD'04. ACM Press, pp 64–69
3. Baralis E, Chiusano S (2004) Essential classification rule sets. ACM Trans Database Syst 29(4):635–674
4. Bastide Y, Taouil R, Pasquier N, Stumme G, Lakhal L (2000) Mining frequent patterns with counting inference. SIGKDD Explor 2(2):66–75
5. Besson J, Pensa RG, Robardet C, Boulicaut J-F (2006) Constraint-based mining of fault-tolerant patterns from boolean data. In: KDID'05 selected and invited revised papers, vol. 3933 of LNCS, Springer, pp 55–71
6. Boley M, Grosskreutz H (2009) Approximating the number of frequent sets in dense data. Knowl Inf Syst 21(1):65–89
7. Bonchi F, Lucchese C (2006) On condensed representations of constrained frequent patterns. Knowl Inf Syst 9(2):180–201
8. Boulicaut J-F, Bykowski A, Rigotti C (2000) Approximation of frequency queries by means of free-sets. In: Proceedings PKDD'00, vol. 1910 of LNCS, Springer, pp 75–85
9. Boulicaut J-F, Bykowski A, Rigotti C (2003) Free-sets: a condensed representation of boolean data for the approximation of frequency queries. Data Mining Knowl Discov 7(1):5–22
10. Brin S, Motwani R, Silverstein C (1997) Beyond market baskets: generalizing association rules to correlations. In: SIGMOD'97. ACM Press, New york, pp 265–276
11. Bringmann B, Nijssen S, Zimmermann A (2009) Pattern based classification: a unifying perspective. In: LeGo'09 worskhop colocated with ECML/PKDD'09
12. Bringmann B, Zimmermann A (2009) One in a million: picking the right patterns. Knowl Inf Syst 18(1):61–81
13. Brodley CE, Utgoff PE (1995) Multivariate decision trees. Mach Learn 19(1):45–77
14. Calders T, Rigotti C, Boulicaut J-F (2005) A survey on condensed representations for frequent sets. In: Constraint-based mining and inductive databases, vol 3848 of LNCS. Springer, Berlin, pp 64–80
15. Cerf L, Gay D, Selmaoui N, Boulicaut J-F (2008) A parameter free associative classifier. In: Proceedings DaWaK'08, vol 5182 of LNCS. Springer, Berlin, pp 238–247
16. Chang C-C, Lin C-J (2001) LIBSVM: a library for support vector machines'. http://www.csie.ntu.edu.tw/~cjlin/libsvm/
17. Cheng H, Yan X, Han J, Hsu C-W (2007) Discriminative frequent pattern analysis for effective classification. In: Proceedings ICDE'07. IEEE Computer Society, Silver Spring, pp 716–725
18. Cheng H, Yu PS, Han J (2006) AC-close: efficiently mining approximate closed itemsets by core pattern recovery. In: ICDM'06. pp 839–844
19. Cheng J, Ke Y, Ng W (2006) δ-tolerance closed frequent itemsets. In: ICDM'06, pp 139–148
20. Crémilleux B, Boulicaut J-F (2002) Simplest rules characterizing classes generated by delta-free sets. In: Proceedings ES'02. Springer, Berlin, pp 33–46
21. Dong G, Li J (1999) Efficient mining of emerging patterns: discovering trends and differences. In: Proceedings KDD'99. ACM Press, New york, pp 43–52

22. Dong G, Zhang X, Wong L, Li J (1999) CAEP: classification by aggregating emerging patterns. In: Proceedings DS'99, vol 1721 of LNCS, Springer, Berlin, pp 30–42
23. El-Manzalawy Y (2005) WLSVM: integrating libsvm into weka environment. http://www.cs.iastate.edu/~yasser/wlsvm/
24. Fayyad UM, Irani KB (1993) Multi-interval discretization of continous-valued attributes for classification learning. In: Proceedings IJCAI'93. Morgan Kaufmann, Los Altos, pp 1022–1027
25. Fürnkranz J (2002) Round robin classification. J Mach Learn Res 2:721–747
26. Ganter B, Stumme G, Wille R (eds) (2005) Formal concept analysis, foundations and applications, vol 3626 of lecture notes in computer science. Springer, Berlin
27. Garriga GC, Kralj P, Lavrac N (2006) Closed sets for labeled data. In: Proceedings PKDD'06. Springer, Berlin, pp 163–174
28. Garriga GC, Kralj P, Lavrac N (2008) Closed sets for labeled data. J Mach Learn Res 9:559–580
29. Gay D, Selmaoui N, Boulicaut J.-F (2007) Pattern-based decision tree construction. In: Proceedings of IEEE ICDIM'07. IEEE Press, New York, pp 291–296
30. Gay D, Selmaoui N, Boulicaut J-F (2008) Feature construction based on closedness properties is not that simple. In: Proceedings PAKDD'08, vol 5012 of LNCS. Springer, Berlin, pp 112–123
31. Gay D, Selmaoui N, Boulicaut J-F (2009) Application-independent feature construction from noisy samples In: Proceedings PAKDD'09, vol 5476 of LNCS. Springer, Berlin, pp 965–972
32. Hébert C, Crémilleux B (2005) Mining delta-strong characterization rules in large SAGE data. In: PKDD'05 discovery challenge on gene expression data
33. Hébert C, Crémilleux B (2006) Optimized rule mining through a unified framework for interestingness measures. In: Proceedings DaWaK'06, vol 4081 of LNCS. Springer, Berlin, pp 238–247
34. John GH, Langley P (1995) Estimating continuous distributions in bayesian classifiers. In: Proceedings UAI'95. Morgan Kaufmann, Los Altos, pp 338–345
35. Kubica J, Moore AW (2003) Probabilistic noise identification and data cleaning. In: Proceedings ICDM'03. IEEE Computer Society, Silver Spring, pp 131–138
36. Li J, Dong G, Ramamohanarao K (2000) Instance-based classification by emerging patterns. In: Proceedings the 4th European conference on principles and practice of knowledge discovery in databases. Springer, Berlin, pp 191–200
37. Li J, Dong G, Ramamohanarao K (2001) 'Making use of the most expressive jumping emerging patterns for classification. Knowl Inf Syst 3(2):131–145
38. Li J, Liu G, Wong L (2007) Mining statistically important equivalence classes and delta-discriminative emerging patterns. In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining KDD'07. ACM Press, New York
39. Li W, Han J, Pei J (2001) CMAR: accurate and efficient classification based on multiple class-association rules. In: Proceedings ICDM'01. IEEE Computer Society, New York, pp 369–376
40. Liu B, Hsu W, Ma Y (1998) Integrating classification and association rule mining. In: Proceedings KDD'98. AAAI Press, pp 80–86
41. Liu G, Li J, Wong L (2007) A new concise representation of frequent itemsets using generators and a positive border. Knowl Inf Syst
42. Miettinen P, Mielikäinen T, Gionis A, Das G, Mannila H (2008) The discrete basis problem. IEEE Trans Knowl Data Eng 20(10):1348–1362
43. Park S-H, Fürnkranz J. (2007) Efficient pairwise classification. In: ECML'07, pp 658–665
44. Pensa RG, Robardet C, Boulicaut J-F (2006) Supporting bi-cluster interpretation in 0/1 data by means of local patterns. Intell Data Anal 10(5):457–472
45. Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann, Los Altos
46. Ramamohanarao K, Fan H (2007) Patterns based classifiers. World Wide Web 10(1):71–83
47. Rebbapragada U, Brodley CE (2007) Class noise mitigation through instance weighting. In: Proceedings ECML'07, vol 4701 of LNCS. Springer, Berlin, pp 708–715
48. Selmaoui N, Leschi C, Gay D, Boulicaut J-F (2006) Feature construction and delta-free sets in 0/1 samples. In: Proceedings DS'06, vol 4265 of LNCS. Springer, Berlin, pp 363–367
49. Soulet A, Crémilleux B, Rioult F (2004) Condensed representation of emerging patterns. In: Proceedings of the 8th Pacific-Asia conference on knowledge discovery in databases, vol 3056 of LNCS, pp 127–132
50. Tan P-N, Steinbach M, Kumar V (2005) Introduction to data mining. Addison-Wesley, Reading
51. Utgoff PE, Brodley CE (1990) An incremental method for finding multivariate splits for decision trees. In: ICML'90, pp 58–65
52. Van Hulse J, Khoshgoftaar TM, Huang H (2007) The pairwise attribute noise detection algorithm. Knowl Inf Syst 11(2):171–190
53. Wang J, Karypis G (2005) HARMONY: efficiently mining the best rules for classification. In: Proceedings SIAM SDM'05, pp 34–43

54. Wang J, Karypis G (2006) On mining instance-centric classification rules. IEEE Trans Knowl Data Eng 18(11):1497–1511
55. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, Los Altos
56. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng AFM, Liu B, Yu PS, Zhou Z-H, Steinbach M, Hand DJ, Steinberg D (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14(1):1–37
57. Yang C, Fayyad UM, Bradley PS (2001) Efficient discovery of error-tolerant frequent itemsets in high dimensions. In: Proceedings KDD'01. ACM Press, New York, pp 194–203
58. Yang Y, Wu X, Zhu X (2004) Dealing with predictive-but-unpredictable attributes in noisy data sources. In: Proceedings PKDD'04, vol 3202 of LNCS. Springer, Berlin, pp 471–483
59. Zhang S, Wu X, Zhang C, Lu J (2008) Computing the minimum-support for mining frequent patterns. Knowl Inf Syst 15(2):233–257
60. Zhang Y, Wu X (2007) Noise modeling with associative corruption rules. In: Proceedings ICDM'07. IEEE Computer Society, New York, pp 733–738
61. Zheng Z (1995) Constructing nominal x-of-n attributes. In: IJCAI'95, pp 1064–1070
62. Zhu X, Wu X (2004) Class noise vs. attribute noise: a quantitative study. Artif Intell Revue 22(3):177–210

Author Biographies

Dominique Gay is currently a post-doctoral researcher in Profiling and Data Mining team at Orange Labs (France Telecom Research and Development). In November 2009, he received a PhD degree in Computer Science from Université de la Nouvelle-Calédonie (Nouméa, New-Caledonia) and Institut National des Sciences Appliquées (Lyon, France). His research themes are about Data Mining and Machine Learning with a special interest for local pattern mining and its use for classification purpose.

Nazha Selmaoui-Folcher is associate professor of computer science at the University of New Caledonia. She is member of PPME laboratory (Multidisciplinary Center of Matter and Environment) since 2008, and she is actually head of computer science team. She is teaching Computer Sciences and Mathematics. She received her PhD degree on Image Analysis. Her research interest is actually on Spatio-Temporal Data Mining and application to environmental sciences. She is a PC member for some data mining conferences.

Jean-François Boulicaut is currently professor of computer science at INSA Lyon. He is the leader of the COMBINING (Modeling and Knowledge Discovery) research group in LIRIS CNRS UMR 5205 and a member of the new INRIA team entitled "Computational Biology and Data Mining" whose goals are to study biological complex systems by means of both modeling/simulation approaches and data mining methods. His own expertise concerns the inductive database framework and the design of unsupervized approaches (e.g., supporting local pattern discovery, co-clustering). He is a member of the Data Mining and Knowledge Discovery journal editorial board since 2006 and has served in the program committees of every major data mining conference.