Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference

Constraint-Based Search of Different Kinds of Discriminative Patterns

Loïc Cerf, João Foscarini, Israel Guerra, Michel Boaventura, Wagner Meira Jr.
Universidade Federal de Minas Gerais
Department of Computer Science, Belo Horizonte, Brazil

Abstract

The state-of-the-art DATA-PEELER algorithm extracts closed patterns in n-ary relations. Because it refines both a lower and an upper bound of the pattern space, DATA-PEELER can, in some circumstances, guarantee that a region of that space does not contain any closed n-set satisfying some relevance constraint. Whenever it happens, such a region is left unexplored and computation saved. This paper shows that some constraints, which DATA-PEELER can efficiently enforce, define useful patterns in the context of a relation with groups of elements in arbitrary dimensions. For instance, it can list the so-called straddling biclusters, which cover at least some given portions of every group. It can discover, as well, closed n-sets that discriminate a group from the others, which are the focus of the experimental section. It shows that DATA-PEELER is highly competitive despite its general enumeration principles and its expressive class of constraints that opens up new applicative perspectives.

Introduction

Given a binary relation, i. e., a set of *objects* described with Boolean *attributes*, the complete extraction of the closed itemsets (Pasquier et al. 1999; Zaki and Hsiao 1999) aims to discover maximal sets of objects sharing the same maximal set of attributes. However, not every closed itemset is worth an interpretation. In practical contexts, the complete collection of all closed itemsets is so large that it is humanly impossible to read them all, if not computationally impossible to list all of them. That is why, since the pioneering works, only a relevant subset of all closed itemsets is searched. This relevance is defined by means of additional constraints every closed itemset must satisfy. Extracting the relevant closed itemsets requires less time if the algorithm is able to identify and prune regions of the pattern space that are guaranteed to not contain any pattern satisfying the additional constraints. This ability depends on both the enumeration principles of the algorithm and the properties of the constraints.

Since the end of the 1990s, several papers have put into focus such properties, hence classes of constraints. Those works are particularly interesting because they move pattern mining closer to what would be an inductive database system (Imielinski and Mannila 1996), i. e., a system where the user can *query* the patterns she is interested in, instead of relying on specific algorithms or, worse, instead of post-processing huge collections of non-specific patterns. This paper defines some generic constraints having many different applications. Those constraints deal with covering, to some extent, some user-defined groups of elements, which can belong to any dimensions of the relation. Here, "dimension" not only means the *objects* or the *attributes* of a binary relation but any of the n dimensions of an n-ary relation. Indeed, the search for the closed itemsets has recently been generalized to n-ary relations. For instance, in a ternary relation where a tuple indicates a customer buying a product during a day, a pattern is a set of customers who bought the same products during the same days. An analyst can focus the search on those 1) involving at least five products that cost more than US$ 10 each (minimal cover of this group of expensive products) and 2) happening at least twice more often during holidays (minimal ratio between the cover of the holidays and the other days).

After a formal definition of the problem, this paper shows that the additional constraints on the patterns are piecewise (anti)-monotone and allow pruning the n-dimensional pattern search space. An efficient verification of those constraints is then described and experiments focusing on the discovery of some discriminative patterns are reported. The related work is detailed at the end of the paper.

Problem Statement

Given $n \in \mathbb{N}$ dimensions of analysis (i. e., n finite sets) $(D_i)_{i=1..n}$, the dataset is a relation $\mathcal{R} \subseteq \times_{i=1}^{n} D_i$, i. e., a set of n-tuples. Table 1 represents such a relation $\mathcal{R}_E \subseteq \{\alpha, \beta, \gamma\} \times \{1, 2, 3, 4\} \times \{A, B, C\}$, hence a ternary relation. In this table, every '1' (resp. '0') at the intersection of three elements stands for the presence (resp. absence) of the related triple in $\mathcal{R}_E$.

A subset of the elements in any of the n dimensions (without loss of generality, they are assumed disjoint) is called *group*. More precisely, a group is a subset of $\cup_{i=1}^{n} D_i$. Table 2 lists three groups, which are defined from the dimensions of $\mathcal{R}_E$. Notice that a group (such as G_3) can involve elements in different dimensions and that two groups can overlap (such as G_1 and G_3) or even be included into each

	A	B	C	A	B	C	A	B	C
1	1	1	1	1	1	1	1	1	0
2	1	1	0	1	0	0	1	1	0
3	0	1	0	0	0	1	1	0	1
4	0	0	1	1	0	1	1	1	1
		α			β			γ	

Table 1: $\mathcal{R}_E \subseteq \{\alpha, \beta, \gamma\} \times \{1, 2, 3, 4\} \times \{A, B, C\}$.

	$\subseteq \{\alpha, \beta, \gamma\}$		$\subseteq \{1, 2, 3, 4\}$		$\subseteq \{A, B, C\}$
G_1	$\emptyset$	$\cup$	$\{1, 2\}$	$\cup$	$\emptyset$
G_2	$\emptyset$	$\cup$	$\{3, 4\}$	$\cup$	$\emptyset$
G_3	$\{\alpha, \beta\}$	$\cup$	$\{2, 3, 4\}$	$\cup$	$\{A, B\}$

Table 2: Three groups defined from the dimensions of $\mathcal{R}_E$.

other (such as G_2 and G_3). This paper deals with the discovery of the closed n-sets involving *at least*, or *at most*, certain user-defined quantities of elements in the groups, as well as defined by the user. More interestingly, it presents a third constraint forcing the ratio of two group covers to be above a given threshold. After recalling the definition of a closed n-set, those three constraints are formally defined.

The *closed n-set* (Cerf et al. 2009) straightforwardly generalizes the famous *closed itemset* (Pasquier et al. 1999; Zaki and Hsiao 1999) to relations defined on (possibly) more than two dimensions:

Definition 1 (Closed n-set) *Given n dimensions of analysis $(D_i)_{i=1..n}$ and a relation $\mathcal{R} \subseteq \times_{i=1}^{n} D_i$, the pattern $(X_1, \ldots, X_n) \in \times_{i=1}^{n} 2^{D_i}$ is a* closed n-set *in $\mathcal{R}$ if and only if:*

Connectedness $\times_{i=1}^{n} X_i \subseteq \mathcal{R}$;

Closedness $\forall (X'_1, \ldots, X'_n) \in \times_{i=1}^{n} D_i$,
$$\begin{cases} \forall i = 1..n, X_i \subseteq X'_i \\ \times_{i=1}^{n} X'_i \subseteq \mathcal{R} \end{cases} \Rightarrow \forall i = 1..n, X_i = X'_i.$$

Example 1 *In $\mathcal{R}_E$ (see Table 1), $(\{\alpha, \gamma\}, \{1, 2\}, \{A, B\})$ is a closed 3-set. $(\{\alpha, \beta\}, \{1, 2\}, \{A, B\})$ is not a closed 3-set because it is not connected ($(\beta, 2, B) \notin \mathcal{R}_E$). $(\{\alpha, \gamma\}, \{1, 2\}, \{A\})$ is not a closed 3-set because it is not closed ($(\{\alpha, \gamma\}, \{1, 2\}, \{A, B\})$ is a strict super-pattern that is connected).*

Given a group $G \subseteq \cup_{i=1}^{n} D_i$, the constraint "covering at least $\mu \in \mathbb{N}$ elements in G" is defined as follows:

Definition 2 (Minimal group cover) $(X_1, \ldots, X_n) \in \times_{i=1}^{n} 2^{D_i}$ *covers at least $\mu \in \mathbb{N}$ elements in G, denoted $\mathcal{C}_{G,\geq,\mu}(X_1, \ldots, X_n)$, if and only if $|(\cup_{i=1}^{n} X_i) \cap G| \geq \mu$.*

"Covering at most $\mu \in \mathbb{N}$ elements in G" is defined in a similar way:

Definition 3 (Maximal group cover) $(X_1, \ldots, X_n) \in \times_{i=1}^{n} 2^{D_i}$ *covers at most $\mu \in \mathbb{N}$ elements in G, denoted $\mathcal{C}_{G,\leq,\mu}(X_1, \ldots, X_n)$, if and only if $|(\cup_{i=1}^{n} X_i) \cap G| \leq \mu$.*

Example 2 *Given the groups in Table 2, the pattern $(\{\alpha, \gamma\}, \{1, 2\}, \{A, B\})$ covers both elements in G_1, does not cover any element in G_2 and covers four elements — α, 2, A and B — in G_3. As a consequence, it* minimally *(resp.* maximally*) covers these three groups if and only if the related minimal (resp. maximal) thresholds are lesser (resp. greater) or equal to, respectively,* 0, 2 *and* 4.

Finally, a minimal ratio $\rho \in \mathbb{R}$ between the covers of two groups G and G' is enforced by the following constraint:

Definition 4 (Minimal cover ratio) *The group cover of $(X_1, \ldots, X_n) \in \times_{i=1}^{n} 2^{D_i}$ exceeds the minimal ratio $\rho \in \mathbb{R}$ between G and G', denoted $\mathcal{C}_{G,G',\rho}(X_1, \ldots, X_n)$, if and only if $\frac{|(\cup_{i=1}^{n} X_i) \cap G|}{|(\cup_{i=1}^{n} X_i) \cap G'|} \geq \rho$.*

By swapping the two groups, this constraint can enforce a *maximal* ratio between their covers.

Example 3 *In our running example, the group cover of the pattern $(\{\alpha, \gamma\}, \{1, 2\}, \{A, B\})$ has a ratio $\frac{0}{2} = 0$ between G_1 and G_2, a ratio $\frac{2}{0} = +\infty$ between G_2 and G_1 and a ratio $\frac{2}{4} = 0.5$ between G_2 and G_3. As a consequence, the related minimal ratio constraints are satisfied if and only if the respective thresholds are, at most, those numbers.*

Given an n-ary relation $\mathcal{R} \subseteq \times_{i=1}^{n} D_i$, a finite set $\mathcal{T}_{\min\,/\,\max}$ of triples in $2^{\cup_{i=1}^{n} D_i} \times \{\geq, \leq\} \times \mathbb{N}$ (i. e., a finite set of *groups* associated with *minimal* or *maximal* cover *thresholds* to respect) and a finite set $\mathcal{T}_{\text{ratio}}$ of triples in $2^{\cup_{i=1}^{n} D_i} \times 2^{\cup_{i=1}^{n} D_i} \times \mathbb{R}$ (i. e., a finite set of *pairs of groups* associated with cover *ratios* to exceed), the problem solved in this paper is the computation of $\{(X_1, \ldots, X_n) \in \times_{i=1}^{n} 2^{D_i} \mid \begin{cases} (X_1, \ldots, X_n) \text{ is a closed } n\text{-set} \\ \wedge_{t \in \mathcal{T}_{\min\,/\,\max} \cup \mathcal{T}_{\text{ratio}}} \mathcal{C}_t(X_1, \ldots, X_n) \end{cases}\}$.

For the case $n = 2$, specific instantiations of this problem are found in the literature. The straddling biclusters (Owens III 2009) are closed patterns significantly straddling across user-defined subsets of one dimension D_i, i. e., every group is a subset of D_i and it is always associated with a *minimal* cover threshold ($\mathcal{T}_{\text{ratio}} = \emptyset$ and the domain of $\mathcal{T}_{\min\,/\,\max}$ is restricted to $2^{D_i} \times \{\geq\} \times \mathbb{N}$). In the context of associative classification, popular relevance criteria, for the selection of the patterns, may be expressed with $\mathcal{C}_{G,G',\rho}$, where G and G' are well chosen subsets of the learning objects D_{objects}:

- the classification rules leading to a class c with a minimal *confidence* (Agrawal, Imielinski, and Swami 1993) $\rho \in \mathbb{R}$ are those satisfying $\mathcal{C}_{G_c,D_{\text{objects}},\rho}$, where $G_c \subseteq D_{\text{objects}}$ designates the objects in the class c;
- the classification rules leading to a class c with a minimal *lift* (also called *interest* or *correlation*) (Brin, Motwani, and Ullman 1997) $\rho \in \mathbb{R}$ satisfy $\mathcal{C}_{G_c,D_{\text{objects}},\frac{\rho \times |G_c|}{|D_{\text{objects}}|}}$;
- the patterns emerging in a class c, with a minimal *growth rate* (Dong and Li 1999) $\rho \in \mathbb{R}$, satisfy $\mathcal{C}_{G_c,D_{\text{objects}} \setminus G_c,\rho}$.

This list is not exhaustive and the constraints $\mathcal{C}_{G,\geq,\mu}$ and $\mathcal{C}_{G,\leq,\mu}$ can complement or substitute $\mathcal{C}_{G,G',\rho}$. For instance, the patterns that are jumping emerging (Li, Dong, and Ramamohanarao 2000) in the class c satisfy $\mathcal{C}_{D_{\text{objects}} \setminus G_c,\leq,0}$ and,

in Cerf et al. (2008), the patterns satisfying $\wedge_{j\neq i}\mathcal{C}_{G_{c_j},\leq,\mu_j} \wedge \mathcal{C}_{G_{c_i},\geq,\mu_i}$ discriminate a class c_i from every other class c_j.

Guided Traversal of the Pattern Space

The problem stated in the previous section can be solved by first extracting all closed n-sets and then filtering those satisfying the constraints. That solution is intractable unless the relation $\mathcal{R}$ is very small. Indeed, there are, in the worst case, $2^{\sum_{i\neq j}|D_i|}$ closed n-sets to be extracted from $\mathcal{R}$ (where j is the index of the largest dimension). To lower the time requirements, the constraints must be enforced during the closed n-set extraction. More precisely, they must guide the search for the valid patterns, i. e., allow to prune regions of the pattern space that need not be traversed because they do not contain any valid closed n-set. Fortunately, the three constraints happen to be *piecewise (anti)-monotone*.

Piecewise (Anti)-Monotonicity

The definition of piecewise (anti)-monotonicity relies on the notion of (anti)-monotonicity per argument:

Definition 5 *Given n dimensions of analysis $(D_i)_{i=1..n}$ and $i = 1..n$, a constraint $\mathcal{C}$ is* (anti)-monotone *w.r.t. the i^{th} argument if and only if* $\forall (X_1,\ldots,X_n) \in \times_{i=1..n} 2^{D_i}$,

$$\begin{cases} \textit{(monotonicity)}\ \forall Y_i \in D_i, X_i \subseteq Y_i \Rightarrow \\ (\mathcal{C}(X_1,\ldots,X_n) \Rightarrow \mathcal{C}(X_1,\ldots,X_{i-1},Y_i,X_{i+1},\ldots,X_n)) \\ \textit{or} \\ \textit{(anti-monotonicity)}\ \forall Y_i \in D_i, X_i \subseteq Y_i \Rightarrow \\ (\mathcal{C}(X_1,\ldots,X_{i-1},Y_i,X_{i+1},\ldots,X_n) \Rightarrow \mathcal{C}(X_1,\ldots,X_n)) \end{cases}$$

Intuitively, a constraint is (anti)-monotone w.r.t. the i^{th} argument if and only if a pattern satisfying the constraint keeps on satisfying it either its i^{th} argument is enlarged or if it is shrunk. A constraint that is (anti)-monotone w.r.t. each *occurrence* of each of its arguments is said piecewise (anti)-monotone:

Definition 6 *A constraint is* piecewise (anti)-monotone *if and only if rewriting it by attributing a separate argument to every occurrence of its variables provides a new constraint that is (anti)-monotone w.r.t. each of its arguments.*

Neither $\mathcal{C}_{G,\geq,\mu}$ nor $\mathcal{C}_{G,\leq,\mu}$ need to be rewritten to be proven piecewise (anti)-monotone. Indeed, each of their variables occurs only once in their expressions (see Definition 2 and 3). $\mathcal{C}_{G,\geq,\mu}$ is monotone w.r.t. all its n arguments. $\mathcal{C}_{G,\leq,\mu}$ is anti-monotone w.r.t. all its n arguments. On the contrary, proving $\mathcal{C}_{G,G',\rho}$ piecewise (anti)-monotone requires rewriting its expression (see Definition 4) with one separate argument per occurrence of its variables $(X_i)_{i=1..n}$:

$$\mathcal{C}'_{G,G',\rho}(X_1,\ldots,X_n,X'_1,\ldots,X'_n) \equiv \frac{|(\cup_{i=1}^n X_i)\cap G|}{|(\cup_{i=1}^n X'_i)\cap G'|} \geq \rho$$

This rewritten constraint $\mathcal{C}'_{G,G',\rho}$ is (anti)-monotone w.r.t. each of its arguments because, assuming it is satisfied, it keeps on being satisfied when any:

- X_i is enlarged (monotonicity w.r.t. the n first arguments);
- X'_i is shrunk (anti-monotonicity w.r.t. the n last arguments).

Pruning the Pattern Space

DATA-PEELER is the state-of-the-art closed n-set extractor. It explores the pattern space by traversing a binary tree whose nodes are associated with two patterns, namely $(L_1,\ldots,L_n)$ and $(U_1,\ldots,U_n)$. Please refer to Cerf et al. (2009) for a detailed presentation of this traversal and why it supports a correct and complete discovery of the closed n-sets. For this paper, the only relevant property is that any closed n-set $(X_1,\ldots,X_n)$ that may be discovered by traversing an enumeration sub-tree is such that $\forall i = 1..n$, $L_i \subseteq X_i \subseteq U_i$, where $(L_1,\ldots,L_n)$ and $(U_1,\ldots,U_n)$ are the patterns associated with the root of the sub-tree. In other terms, $(L_1,\ldots,L_n)$ and $(U_1,\ldots,U_n)$ are a lower and an upper bound of the pattern space that would be explored "below" the current node. The conditional tense applies because, thanks to those bounds, DATA-PEELER efficiently verifies whether that pattern space can possibly contain a closed n-set satisfying chosen piecewise (anti)-monotone constraints and, if it does not, this pattern space is not explored, i. e., the binary tree is pruned at the current node.

In the context of this paper, here is the predicate that DATA-PEELER evaluates to decide whether it is safe to prune the binary tree at the current node:

$$\begin{aligned} &(\exists (G,\geq,\mu) \in \mathcal{T}_{\min\,/\,\max} \mid \neg\mathcal{C}_{G,\geq,\mu}(U_1,\ldots,U_n)) \\ \vee&(\exists (G,\leq,\mu) \in \mathcal{T}_{\min\,/\,\max} \mid \neg\mathcal{C}_{G,\leq,\mu}(L_1,\ldots,L_n)) \\ \vee&(\exists (G,G',\rho) \in \mathcal{T}_{\text{ratio}} \mid \neg\mathcal{C}'_{G,G',\rho}(U_1,\ldots,U_n,L_1,\ldots,L_n)) \end{aligned}$$

In this predicate, U_i replaces the monotone occurrences of the variable X_i and L_i replaces the anti-monotone occurrences of this same variable. If the predicate is satisfied then, the piecewise (anti)-monotonicity of the constraints guarantees that for all pattern $(X_1,\ldots,X_n) \in \times_{i=1}^n 2^{D_i}$ such that $\forall i = 1..n$, $L_i \subseteq X_i \subseteq U_i$, we have:

$$\begin{aligned} &(\exists (G,\geq,\mu) \in \mathcal{T}_{\min\,/\,\max} \mid \neg\mathcal{C}_{G,\geq,\mu}(X_1,\ldots,X_n)) \\ \vee&(\exists (G,\leq,\mu) \in \mathcal{T}_{\min\,/\,\max} \mid \neg\mathcal{C}_{G,\leq,\mu}(X_1,\ldots,X_n)) \\ \vee&(\exists (G,G',\rho) \in \mathcal{T}_{\text{ratio}} \mid \neg\mathcal{C}_{G,G',\rho}(X_1,\ldots,X_n)) \end{aligned}$$

In other terms, any pattern, which would be considered in the sub-tree rooted by the current node, is violating at least one constraint and there is no need to traverse this sub-tree.

Implementation

The predicate, stated at the end of the previous section, is evaluated after the construction of every enumeration node to decide whether the sub-tree it roots is safely pruned. A naive evaluation of the predicate would simply stick to the Definitions 2, 3 and 4, i. e., would intersect the elements in each group with those in the (lower or upper) bound. Assuming the elements are maintained ordered, the time complexity of this naive evaluation would be linear in the total number of elements (in all groups and in the two bounds). However, by simply understanding that DATA-PEELER refines a lower and an upper bound of the pattern space, one can easily implement a faster evaluation of the predicate.

Indeed, from an enumeration node to one of its two children, DATA-PEELER adds some elements to the lower bound

Dataset	$\lvert D_{\text{objects}}\rvert$	$\lvert G_{\min}\rvert$	$\lvert D_{\text{attributes}}\rvert$	$\lvert\mathcal{R}\rvert$	$\lvert\mathcal{R}\rvert/\lvert D_1 \times D_2\rvert$
adult	48842	11687	14	97684	0.1428
cylBands	540	228	34	1080	0.0588
letRecog	20000	734	15	40000	0.1333
penDigits	10992	1055	16	21984	0.1250
soybean	683	8	35	1366	0.0571
waveform	5000	1647	21	10000	0.0950

Table 3: Characteristics of the datasets: number of objects, number of objects in the smallest class, number of attributes, number of tuples and density (in this order).

and removes some elements from the upper bound. This incremental (resp. decremental) computation of the lower (resp. upper) bound allows a faster evaluation of the predicate. It is only about maintaining updated, along the pattern space traversal, the quantity $l_G = |\cup_{i=1}^{n} L_i \cap G|$ (resp. $u_G = |\cup_{i=1}^{n} U_i \cap G|$) for every group that must be maximally (resp. minimally) covered and for every group appearing at the denominator (resp. numerator) of the cover ratios. In this way, these cardinalities can be compared, in constant time, to the maximal (resp. minimal) cover thresholds and the quotients $\frac{u_G}{l_{G'}}$ can be compared, as well in constant time, to the minimal cover ratio ρ of the constraints $\mathcal{C}_{G,G',\rho}$. Given the set $\cup_{i=1..n} L_i^{\text{child}} \setminus L_i^{\text{parent}}$ (resp. $\cup_{i=1..n} U_i^{\text{parent}} \setminus U_i^{\text{child}}$) of the elements that are added to (resp. removed from) the lower (resp. upper) bound, when going from the parent node to the child node, we have $l_G^{\text{child}} = l_G^{\text{parent}} + |\{e \in \cup_{i=1..n} L_i^{\text{child}} \setminus L_i^{\text{parent}} \mid e \in G\}|$ and $u_G^{\text{child}} = u_G^{\text{parent}} - |\{e \in \cup_{i=1..n} U_i^{\text{parent}} \setminus U_i^{\text{child}} \mid e \in G\}|$.

Testing whether an element e is in a group G requires a constant time if a bitset represents the group[1]. Overall, the time complexity of this evaluation of the predicate is linear in the number of groups $|\mathcal{T}_{\min\,/\,\max}| + 2|\mathcal{T}_{\text{ratio}}|$ multiplied by the number $|\cup_{i=1..n} (L_i^{\text{child}} \setminus L_i^{\text{parent}}) \cup (U_i^{\text{parent}} \setminus U_i^{\text{child}})|$ of elements that were added to the lower bound or removed from the upper bound. In practice, this complexity is far below that of the naive evaluation.

Experimental Study

Our C++ code was compiled with GCC 4.5.3 with the O3 optimizations. This section reports experiments performed on a GNU/Linux™ system running on top of a core cadenced at 3.4 GHz.

Extracting Confident Classification Rules

This section reports running times for the extraction of classification rules under the minimal confidence constraint (Agrawal, Imielinski, and Swami 1993). The most demanding normalized and discretized datasets from Coenen (2003) (but connect4 whose patterns fill the whole disk in some of the considered settings) are used. Their characteristics are listed in Table 3. We focus on the extraction of the rules predicting the smallest class, which is usually considered the hardest to characterize. As a consequence, DATA-PEELER's enumeration is constrained by $\mathcal{C}_{G_{\min},D_{\text{objects}},\rho}$, where $G_{\min}$ is the smallest class, D_{objects} the set of all learning objects (including those in the smallest class) and ρ the minimal confidence that varies between 0 and 1. Figure 1 presents DATA-PEELER's running times as well as those obtained with the post-processing approach, which consists in first listing all closed itemsets (the latest version of LCM (Uno, Kiyomi, and Arimura 2005) was used) and then filtering those satisfying the minimal confidence (with a homemade C++ program).

On a given dataset, the running time of LCM and its post-processing is constant. Such result is expected since the same closed itemsets are mined and post-processed in all settings. DATA-PEELER's extractions are faster when the minimal confidence is higher. For instance, with the letRecog dataset, it is divided by 3.5 between $\rho = 0$ and $\rho = 1$. This reflects the more pruning enabled by a stronger constraint. DATA-PEELER can run up to 3 orders of magnitude faster than LCM followed by the post-processing. For instance, with the soybean dataset and a 1-confidence, LCM and the post-processing extract the classification rules in 90 seconds, whereas DATA-PEELER requires only 0.1 second.

Extracting Patterns of Influence in Twitter

To the best of our knowledge, DATA-PEELER is the fastest closed n-set extractor when $n \geq 3$. That explains why our approach was so far compared with LCM + post-processing in the restricted case $n = 2$. However, $\mathcal{C}_{G,\geq,\mu}$, $\mathcal{C}_{G,\leq,\mu}$ and $\mathcal{C}_{G,G',\rho}$ are useful in the broader context of an n-ary relation. The dataset, used in this section, is 3-dimensional. It was constructed from messages published on Twitter, a famous microblogging platform. Messages about the Brazilian soccer championship were collected from the 19th week of 2012 to the 30th week of this same year. They were classified w.r.t. to the mentioned team(s) (supervised classification method, which is out of the scope of this paper). A user is considered influential (Kwak et al. 2010) about a given team during a given week if at least one message she wrote about the team during this week was retweeted (i. e., other users "repeated" the message). A ternary relation lists the 790,466 such triples (user,team,week). Overall, they involve 307,707 users, all 20 teams in the *Série A* and 12 weeks.

Let us assume the analyst wants to discover famous users who are frequently influential when they write about the São Paulo's teams but not influential w.r.t. teams outside São Paulo. For that purpose, three groups are defined:

[1]The bitsets representing the groups are constructed after the relation $\mathcal{R}$ is read (so that all elements are known and attributed an id, its index in the bitsets). They are never modified afterward.

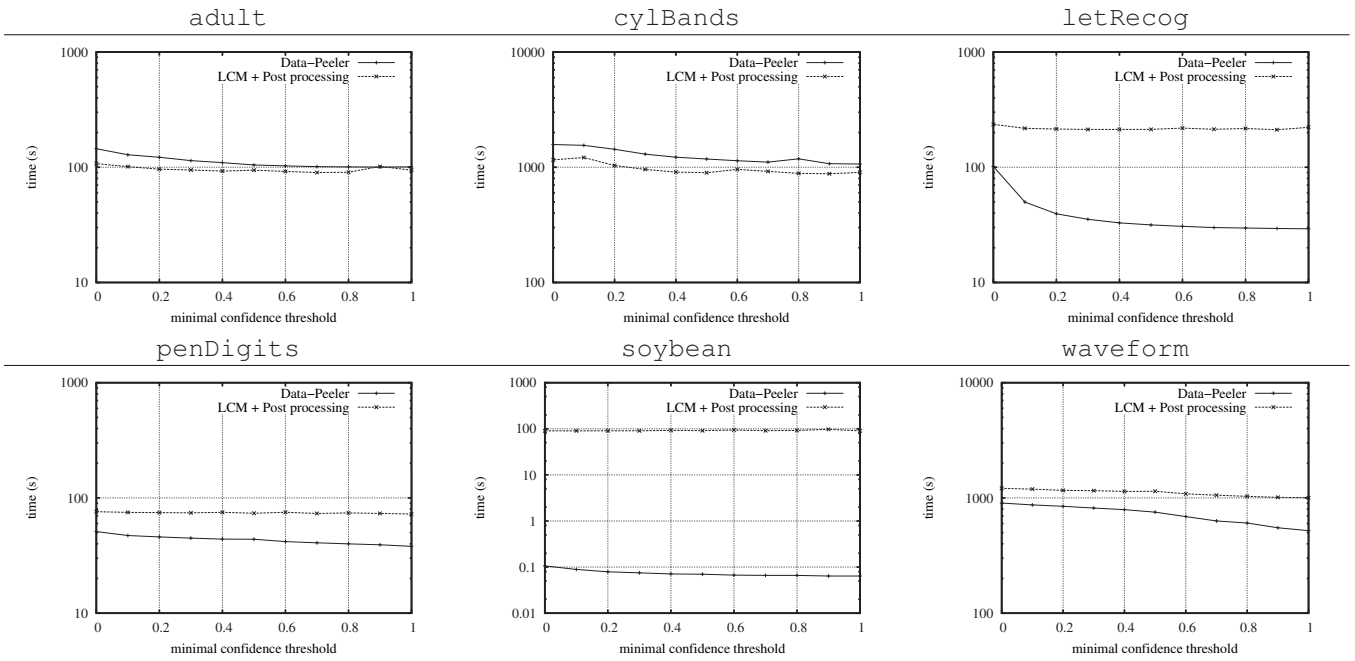

Figure 1: Running times for the extraction of the confident rules concluding on the smallest class.

G_1 contains the six teams from São Paulo;

G_2 contains the remaining fourteen teams;

G_3 contains the "famous" users defined as the top 10,000 users in numbers of retweets during the studied period of time (use of the original numerical data).

DATA-PEELER is used to extract patterns involving at least ten weeks (classical frequency constraint that forces the users to be retweeted almost every week), 60 users and the following conjunction of group cover constraints:

$$\mathcal{C}_{G_1,\geq,2} \wedge \mathcal{C}_{G_2,\leq,0} \wedge \mathcal{C}_{G_3,\geq,60}$$

The conjunction of the first two constraints actually defines a jumping emerging pattern (Li, Dong, and Ramamohanarao 2000) covering at least two teams from São Paulo. DATA-PEELER discovers two closed 3-sets after 12 minutes and 53 seconds. In this experiment, the group cover constraints do not filter any more pattern, i. e., the same two closed 3-sets are discovered with the sole frequency constraints. However, because the group cover constraints more finely specify those patterns, the running time is lowered. Indeed, without the group cover constraints, the extraction lasts 14 minutes.

The two closed 3-sets, which are discovered, involve the same two teams: Corinthians and Palmeiras. They are the two most popular ones from São Paulo. The users involved in the patterns are indeed "correlated" with Corinthians and Palmeiras. To reach this conclusion, we simply count the numbers of unique users whose names include a string that is characteristic of the 20 teams (e. g., "corin", "timao", "fiel" for Corinthians or "sant", "sfc", "peixe" for Santos, another team from São Paulo). In this way, we find out 29 users supporting Palmeiras, 24 users supporting Corinthians and only 2 users supporting Santos, a third team from São Paulo. The remaining 161 user names do not relate to any team.

Related Work

Piecewise (anti)-monotone constraints support the definition of complex patterns. This class of constraints includes (but is not restricted to) any Boolean expression of monotone (Grahne, Lakshmanan, and Wang 2000) and anti-monotone (Ng et al. 1998) constraints. It includes as well constraints that are neither succinct (Ng et al. 1998) nor convertible (Pei and Han 2000) nor loose anti-monotone (Bonchi and Lucchese 2005). $\mathcal{C}_{G,G',\rho}$ is such a constraint. As mentioned earlier, it supports the definition of various patterns that are the bases of associative classifiers and have been usually listed by post-processing huge collections of non-specific patterns, hence a scalability issue. Thanks to DATA-PEELER's ability to prune the pattern space with any piecewise (anti)-monotone constraint, those discriminative patterns can now be efficiently discovered in more general contexts (n-ary relations with groups of objects in *any* dimensions). This opens up many perspectives such as the classification of n-dimensional objects, the rapid development of associative classifiers relying on new patterns that can be expressed with $\mathcal{C}_{G,\geq,\mu}$, $\mathcal{C}_{G,\leq,\mu}$ and $\mathcal{C}_{G,G',\rho}$, and, with little more effort, the efficient enforcement of any other piecewise (anti)-monotone constraint.

Like Cerf et al. (2009), this paper defines the piecewise (anti)-monotonicity in a "divisive" way (see Definition 5): they are constraints which are either monotone or anti-monotone w.r.t. each *occurrence* of a variable in their expressions. Soulet and Crémilleux (2005) have independently proposed a "constructive" definition of the same class of

constraints: those constraints are recursively defined from arbitrary primitives that either increase or decrease w.r.t. each of their arguments. The proposed algorithm is, however, restricted to binary relations. To the best of our knowledge, Soulet and Crémilleux (2009) have written the most comprehensive study of the piecewise (anti)-monotone constraints and their ability to prune the pattern space.

Guns, Nijssen, and Raedt (2011) rely on constraint programming to mine patterns under constraints. This allows declaring a wide range of constraints. As shown more thoroughly by Nijssen, Guns, and Raedt (2009), that includes constraints that define various discriminative patterns. Nevertheless, the expressiveness of this framework remains unclear and all constraints, in those articles, are piecewise (anti)-monotone. More importantly, general purpose solvers somehow are too general, i. e., they do not specifically suit the pattern mining task. As a consequence, scalability becomes problematic. In particular, the extraction of the frequent closed itemsets is about 100 times slower than with LCM (Guns, Nijssen, and Raedt 2011). On the contrary, DATA-PEELER is competitive even when the relation is binary and the sole frequency constraint is enforced.

Conclusion

All along the traversal of the n-dimensional pattern space, DATA-PEELER knows the lower and the upper bound of the patterns it could recursively enumerate. The conditional tense applies: if the relevant patterns are to satisfy a piecewise (anti)-monotone constraint, the recursive search may be aborted with the guarantee that no such pattern is missed. Three simple piecewise (anti)-monotone constraints can be combined in various ways to define different discriminative patterns. In this way, DATA-PEELER is a generic yet competitive solution to the discovery of all those patterns.

Acknowledgments

This work was supported by the CNPq, FAPEMIG and InWeb.

References

Agrawal, R.; Imielinski, T.; and Swami, A. N. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 SIGMOD International Conference on Management of Data*, 207–216. Washington, D.C., USA: ACM Press.

Bonchi, F., and Lucchese, C. 2005. Pushing tougher constraints in frequent pattern mining. In *Proceedings of the Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 114–124. Hanoi, Vietnam: Springer.

Brin, S.; Motwani, R.; and Ullman, J. D. 1997. Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the 1997 SIGMOD International Conference on Management of Data*, 265–276. Tucson, Arizona: ACM Press.

Cerf, L.; Gay, D.; Selmaoui, N.; and Boulicaut, J.-F. 2008. A parameter-free associative classification method. In *Proceedings of the Tenth International Conference on Data Warehousing and Knowledge Discovery*, 293–304. Turin, Italy: Springer.

Cerf, L.; Besson, J.; Robardet, C.; and Boulicaut, J.-F. 2009. Closed patterns meet n-ary relations. *ACM Transactions on Knowledge Discovery from Data* 3(1):1–36.

Coenen, F. 2003. The lucs-kdd discretised/normalised arm and carm data library. www.csc.liv.ac.uk/ frans/KDD/Software/LUCS-KDD-DN/.

Dong, G., and Li, J. 1999. Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the Fifth SIGKDD International Conference on Knowledge Discovery and Data Mining*, 43–52. San Diego, California, USA: ACM Press.

Grahne, G.; Lakshmanan, L. V. S.; and Wang, X. 2000. Efficient mining of constrained correlated sets. In *Proceedings of the Sixteenth International Conference on Data Engineering*, 512–521. San Diego, California, USA: IEEE Computer Society.

Guns, T.; Nijssen, S.; and Raedt, L. D. 2011. Itemset mining: A constraint programming perspective. *Artificial Intelligence* 175(12–13):1951–1983.

Imielinski, T., and Mannila, H. 1996. A database perspective on knowledge discovery. *Communications of the ACM* 39(11):58–64.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In *Proceedings of the Nineteenth International Conference on World Wide Web*, 591–600. Raleigh, North Carolina, USA: ACM Press.

Li, J.; Dong, G.; and Ramamohanarao, K. 2000. Making use of the most expressive jumping emerging patterns for classification. In *Proceedings of the Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 220–232. Kyoto, Japan: Springer.

Ng, R. T.; Lakshmanan, L. V. S.; Han, J.; and Pang, A. 1998. Exploratory mining and pruning optimizations of constrained association rules. In *Proceedings of the 1998 SIGMOD International Conference on Management of Data*, 13–24. Seattle, Washinton, USA: ACM Press.

Nijssen, S.; Guns, T.; and Raedt, L. D. 2009. Correlated itemset mining in ROC space: a constraint programming approach. In *Proceedings of the Fifteenth SIGKDD International Conference on Knowledge Discovery and Data Mining*, 647–656. Paris, France: ACM Press.

Owens III, C. C. 2009. Mining truth tables and straddling biclusters in binary datasets. Master's thesis, Faculty of the Virginia Polytechnic Institute and State University.

Pasquier, N.; Bastide, Y.; Taouil, R.; and Lakhal, L. 1999. Efficient mining of association rules using closed itemset lattices. *Information Systems* 24(1):25–46.

Pei, J., and Han, J. 2000. Can we push more constraints into frequent pattern mining? In *Proceedings of the Sixth SIGKDD International Conference on Knowledge Discovery and Data Mining*, 350–354. Boston, Massachusetts, USA: ACM Press.

Soulet, A., and Crémilleux, B. 2005. Exploiting virtual patterns for automatically pruning the search space. In *Proceedings of the Fourth International Workshop on Knowledge Discovery in Inductive Databases*, 202–221. Porto, Portugal: Springer.

Soulet, A., and Crémilleux, B. 2009. Mining constraint-based patterns using automatic relaxation. *Intelligent Data Analysis* 13(1):109–133.

Uno, T.; Kiyomi, M.; and Arimura, H. 2005. LCM ver.3: Collaboration of array, bitmap and prefix tree for frequent itemset mining. In *Proceedings of the First International Workshop on Open Source Data Mining*, 77–86. Chicago, Illinois, USA: ACM Press.

Zaki, M. J., and Hsiao, C.-J. 1999. CHARM: An efficient algorithm for closed association rule mining. Technical Report 99-10, Computer Science Department, Rensselaer Polytechnic Institute, Troy NY 12180.