Mining Formal Concepts with a Bounded Number of Exceptions from Transactional Data

Jérémy Besson[1,2], Céline Robardet[3], and Jean-François Boulicaut[1]

[1] INSA Lyon, LIRIS CNRS FRE 2672, F-69621 Villeurbanne cedex, France
[2] UMR INRA/INSERM 1235, F-69372 Lyon cedex 08, France
[3] INSA Lyon, PRISMA, F-69621 Villeurbanne cedex, France
{Jeremy.Besson, Celine.Robardet, Jean-Francois.Boulicaut}@insa-lyon.fr

Abstract. We are designing new data mining techniques on boolean contexts to identify a priori interesting bi-sets (i.e., sets of objects or transactions associated to sets of attributes or items). A typical important case concerns formal concept mining (i.e., maximal rectangles of true values or associated closed sets by means of the so-called Galois connection). It has been applied with some success to, e.g., gene expression data analysis where objects denote biological situations and attributes denote gene expression properties. However in such real-life application domains, it turns out that the Galois association is a too strong one when considering intrinsically noisy data. It is clear that strong associations that would however accept a bounded number of exceptions would be extremely useful. We study the new pattern domain of α/β concepts, i.e., consistent maximal bi-sets with less than α false values per row and less than β false values per column. We provide a complete algorithm that computes all the α/β concepts based on the generation of concept unions pruned thanks to anti-monotonic constraints. An experimental validation on synthetic data is given. It illustrates that more relevant associations can be discovered in noisy data. We also discuss a practical application in molecular biology that illustrates an incomplete but quite useful extraction when all the concepts that are needed beforehand can not be discovered.

1 Introduction

One of the most popular data mining techniques concerns transactional data analysis by means of set patterns. Transactional data can be represented as boolean matrices. The lines denotes transactions or objects and the columns are boolean attributes that enable to record item occurrences within transactions or properties of objects. For instance, in the toy example $\mathbf{r}_1$ from Figure 1, object o_2 satisfies properties i_1 and i_2 or, alternatively, transaction o_2 contains items i_1 and i_2. Many application domains can lead to such boolean contexts. For instance, beside the classical basket analysis problem where transactions denote the items purchased by some customers, we made many experiments on boolean gene expression data sets that encode gene expression properties in

B. Goethals and A. Siebes (Eds.): KDID 2004, LNCS 3377, pp. 33–45, 2005.

	Items		
	i_1	i_2	i_3
o_1	1	1	1
o_2	1	1	0
o_3	1	0	1
o_4	1	0	0
o_5	0	1	0

Fig. 1. A boolean context $\mathbf{r}_1$

some biological situations (see, e.g., [4]). In this kind of application, the raw data is a collection of numerical values that quantify the activity of each gene in each studied situation. Gene expression properties, for instance over-expression, are then computed by means of discretization techniques (see, e.g., [1, 14]). For example, given $\mathbf{r}_1$, we might say that all the studied genes are considered over-expressed in situation o_1.

Given eventually huge transactional data sets, hundreds of research papers have considered the efficient computation of a priori interesting association rules from the so-called frequent sets of attributes. Also, the multiple uses of (frequent) closed sets of transactions and/or attributes have been studied a lot. In this paper, we consider bi-set mining from transactional data. More precisely, we want to compute sets of objects T and sets of attributes G that are strongly associated within the data. An interesting case concerns formal concept discovery, i.e., the computation of maximal rectangles of true values [20]. For instance, in $\mathbf{r}_1$, $(\{o_1, o_2\}, \{i_1, i_2\})$ is a formal concept or concept for short. In boolean gene expression data sets, concepts can be considered as putative transcription modules, i.e., maximal sets of genes that are co-regulated associated to the maximal set of situations in which they are co-regulated. Their discovery is an important step towards the understanding of gene regulation networks. It is the major application domain which motivates our research.

Collections of concepts can be used, e.g., for conceptual clustering or as condensed representations for association rules. Efficient algorithms enable to compute concepts [8, 2, 11]. When the extraction task is too hard, it is also possible to compute concepts under constraints. It can be based on (frequent) closed set computation (see, e.g., [12, 5, 13, 21, 6, 7, 16] and [10] for a recent survey). It is also possible to use an algorithm that directly mine concepts under constraints on both set components [3].

The aim of concept extraction is to identify objects and properties which are strongly associated. Within a concept, we have a maximal set of objects (i.e., a closed set) which are in relation with all the elements of a maximal set of properties and vice versa. This degree of association is often too strong in real-life data. This is typical in life sciences where we can not avoid error of measurement or when discretization methods are used and can easily lead to some wrong values. Indeed, once a discretization threshold has been computed (say 34.5) for deciding about the over-expression of a given gene, assigning false (and thus not over-expression) for a situation whose raw expression value is 34 might be or not an error. What is clear, is that concepts that would accept exceptions could be extremely useful. Assume that in a boolean context, we have a bi-set (T, G) (with, e.g., $|T| = 12$ and $|G| = 25$) such that each property from G is not shared by at most one object from T and each object from T does not have at most two properties from G. Our thesis is that it is extremely useful to extract such a bi-set for further post-processing by data owners. Indeed the presence of erroneous false values in the data set leads to the multiplication of concepts from which it might be hard to identify the relevant associations. As an illustration, in Figure 1, the bi-set $(\{o_1, o_2, o_3\}, \{i_1, i_2, i_3\})$ is not a concept but has at most 1 false value per row and at most 1 false value per column. It appears to be the union of 4 concepts which are $(\{o_1\}, \{i_1, i_2, i_3\})$, $(\{o_1, o_2, o_3\}, \{i_1\})$, $(\{o_1, o_2\}, \{i_1, i_2\})$, and $(\{o_1, o_3\}, \{i_1, i_3\})$.

Therefore, the contribution of this paper is to propose a new kind of patterns called the α/β concepts, i.e., concepts with exceptions or, more precisely, maximal consistent bi-sets of true values with a bounded number of false values per row (α threshold) and per column (β threshold). Therefore, we specify the desired patterns within a constraint-based mining framework. The constraint $\mathcal{C}_{\alpha\beta}$ is used to enforce a bounded number of exceptions. The consistency constraint denoted $\mathcal{C}_{cons}$ is important: only relevant patterns such that there is no row (resp. column) outside the bi-set which is identical to an inside one w.r.t. the bi-set columns (resp. rows) have to be mined. Finally, we also use maximality constraints (denoted $\mathcal{C}_{max}$) w.r.t. the collections specified by the other constraints and our specialization relation on bi-sets. We studied how to compute α/β concepts. This is indeed a difficult problem since we loose the Galois connection properties in this new setting. Our main theoretical result concerns the formalization of a constraint-based mining framework that can be used for computing every α/β concept. For that purpose, we start by computing every concept and then we perform unions of concepts while "pushing" the constraints $\mathcal{C}_{\alpha\beta}$, $\mathcal{C}_{cons}$, and $\mathcal{C}_{max}$ to reduce the search space. Doing so, the complete collection of α/β concepts can be computed. We provide two experimental validations. First, we consider a synthetic data set. This data set consists of some formal concepts and uniform random noise. We show that α/β concept mining enables to discover the original associations (i.e., the concepts that were existing before noise introduction) provided that the noise is not too important. Then, we discuss a practical application in molecular biology. It illustrates an incomplete but quite useful extraction when all the α/β concepts can not be discovered: instead of computing the whole collection of α/β concepts we compute a subset of them obtained from large enough concept unions. By this application we demonstrate that large α/β concepts can be computed that contain a rather small number of exceptions.

The paper is organized as follows. In Section 2, we provide the needed definitions and the formalization of the α/β concept mining task. Section 3 sketches the algorithm and discusses its properties. Section 4 concerns the experimental validation of our approach. Finally, Section 5 is a short conclusion.

2 Formalizing α/β Concept Mining

Let $\mathcal{O}$ denotes a set of objects and $\mathcal{P}$ denotes a set of properties. The transactional data or boolean context is $\mathbf{r} \subseteq \mathcal{O} \times \mathcal{P}$. $(o_i, i_j) \in \mathbf{r}$ denotes that property j holds for object i. A bi-set is an element of $\mathcal{L} = \mathcal{L}_\mathcal{O} \times \mathcal{L}_\mathcal{P}$ where $\mathcal{L}_\mathcal{O} = 2^\mathcal{O}$ and $\mathcal{L}_\mathcal{P} = 2^\mathcal{P}$.

Definition 1 (1-Rectangle). *A bi-set (T, G) is a 1-rectangle in $\mathbf{r}$ iff $\forall t \in T$ and $\forall g \in G$ then $(t, g) \in \mathbf{r}$. We say that it satisfies constraint $\mathcal{C}_{1R}(T, G)$. When a bi-set (T, G) is not a 1-rectangle, we say that it contains 0 values.*

Definition 2 (Concept). *A bi-set (T, G) is a concept in $\mathbf{r}$ iff (T, G) is a 1-rectangle and $\forall T' \subseteq \mathcal{O} \backslash T$, $(T \cup T', G)$ is not a 1-rectangle and $\forall G' \subseteq \mathcal{P} \backslash G$, $(T, G \cup G')$ is not a 1-rectangle. A concept (T, G) is thus a maximal 1-rectangle.*

Example 1. *$(\{o_1\}, \{i_1, i_3\})$ is a 1-rectangle in $\mathbf{r}_1$ but it is not a concept. An example of a concept in $\mathbf{r}_1$ is $(\{o_1, o_3\}, \{i_1, i_3\})$.*

By construction, concepts are built on two so-called closed sets that are associated by the Galois connection.

Definition 3 (Galois Connection [20]). *If $T \subseteq \mathcal{O}$ and $G \subseteq \mathcal{P}$, assume $\phi(T, \mathbf{r}) = \{i \in \mathcal{P} \mid \forall o \in T,\ (o, i) \in \mathbf{r}\}$ and $\psi(G, \mathbf{r}) = \{o \in \mathcal{O} \mid \forall i \in G,\ (o, i) \in \mathbf{r}\}$. ϕ provides the set of items that are common to a set of objects and ψ provides the set of objects that share a set of items. (ϕ, ψ) is the so-called Galois connection between $\mathcal{O}$ and $\mathcal{P}$. We use the classical notations $h = \phi \circ \psi$ and $h' = \psi \circ \phi$ to denote the Galois closure operators. A set $T \subseteq \mathcal{O}$ (resp. $G \subseteq \mathcal{P}$) is said closed iff $T = h'(T)$ (resp. $G = h(G)$).*

An important property of the Galois connection is that each closed set on one dimension is associated to a unique closed set on the other dimension. It explains why any algorithm that computes closed sets can be used for concept extraction (see, e.g., [16] for a discussion when using a frequent closed set computation algorithm in the context of gene expression data analysis).

Example 2. *$(\{o_1, o_2\}, \{i_1, i_2\})$ is a concept in $\mathbf{r}_1$. We have $h(\{i_1, i_2\}) = \{i_1, i_2\}$, $h'(\{o_1, o_2\}) = \{o_1, o_2\}$, $\phi(\{o_1, o_2\}) = \{i_1, i_2\}$, and $\psi(\{i_1, i_2\}) = \{o_1, o_2\}$.*

Many algorithms like AC-Miner[1] [6], CHARM [21] and CLOSET+ [19] extract frequent closed sets and thus very easily concepts under a minimal frequency constraint on the set of objects by an application of one of the Galois operators. This user-defined minimal frequency constraint enables to optimize the extraction tasks in dense and/or highly correlated data sets: both the search

[1] Even though this algorithm has been designed for the extraction of frequent δ-free sets, we often use the formal property which states that every frequent closed set is the closure of a 0-free set. In other terms, a straightforward postprocessing on AC-Miner output can provide every frequent closed set.

space and the solution space can be reduced. In practice, we can have however too large or too dense matrices (see, e.g., the case of some biological contexts in Section 4.2) such that only very high minimal frequency thresholds can lead to tractable computations. Assuming a standard boolean context, it means that only bi-sets composed of few items and many objects can be extracted whereas we would like to enforce other constraints. To overcome this problem, we have proposed in [3] the algorithm D-Miner which enables to extract formal concepts while "pushing" other meaningful constraints.

Definition 4 (Meaningful Constraints on Concepts).
Minimal size constraints: a concept (T, G) *satisfies the constraint* $\mathcal{C}_{ms}(\mathbf{r}, \sigma_1, \sigma_2, (T, G))$ *iff* $|T| \geq \sigma_1$ *and* $|G| \geq \sigma_2$.
Syntactical constraints: a concept (T, G) *satisfies the constraint* $\mathcal{C}_{Inclusion}(\mathbf{r}, X, Y, (T, G))$ *iff* $X \subseteq T$ *and* $Y \subseteq G$.
Minimal area constraint: a concept (T, G) *satisfies the constraint* $\mathcal{C}_{area}(\mathbf{r}, \sigma, (T, G))$ *iff* $|T| \times |G| \geq \sigma$.

More precisely, D-Miner extract efficiently formal concepts which moreover satisfy some monotonic constraints w.r.t. the following specialization relation.

Definition 5 (Specialization Relation). *Our specialization relation on bi-sets from* $\mathcal{L}_{\mathcal{O}} \times \mathcal{L}_{\mathcal{P}}$ *is defined by* $(T_1, G_1) \preceq (T_2, G_2)$ *iff* $T_1 \subseteq T_2$ *and* $G_1 \subseteq G_2$. *As usual,* $\prec$ *is used to denote strict specialization (i.e., using* $\subset$ *instead of* $\subseteq$*).*

Definition 6 (Monotonic Constraints on Bi-sets). *Given* $\mathcal{L}$ *a collection of bi-sets, a constraint* $\mathcal{C}$ *is said monotonic w.r.t.* $\subseteq$ *iff* $\forall \alpha$, $\beta \in \mathcal{L}$ *such that* $\alpha \subseteq \beta$, $\mathcal{C}(\alpha) \Rightarrow \mathcal{C}(\beta)$.

Example 3. *The three previously defined constraints are examples of monotonic constraints w.r.t. our specialization relation. The concepts (T,G) satisfying* $\mathcal{C}_{ms}(\mathbf{r}_1, 2, 2, (T, G))$ *are* $(\{o_1, o_2\}, \{i_1, i_2\})$ *and* $(\{o_1, o_3\}, \{i_1, i_3\})$. *The concepts (T,G) satisfying* $\mathcal{C}_{area}(\mathbf{r}_1, 4, (T, G))$ *are* $(\{o_1, o_2\}, \{i_1, i_2\})$, $(\{o_1, o_3\}, \{i_1, i_3\})$ *and* $(\{o_1, o_2, o_3, o_4\}, \{i_1\})$.

A concept (T, G) is such that all its items and objects are in relation. Thus, the absence of relation between an item g and an object t leads to two concepts, one with g and without t, and another one with t and without g. D-Miner is based on this observation and it builds simultaneously the closed sets of objects and items starting from the bi-set corresponding to the whole sets of items and objects, recursively cutting it using 0 values [3].

Notice that pushing the monotonic constraint $\mathcal{C}_{area}$ within D-Miner is a solution to one of the problems addressed in [9].

The aim of concept extraction is to gather properties and objects which are strongly associated. On another hand, we already motivated the interest of relaxing the maximal 1-rectangle constraint. A simple idea is to consider all the maximal bi-sets with less than α false values per row and less than β false values per column.

Definition 7 ($\alpha\beta$-Constraint). *A bi-set (T, G) satisfies $\mathcal{C}_{\alpha\beta}$ in $\mathbf{r}$ iff*
$\forall o \in T, |\{i \in G \text{ such that } (o, i) \notin \mathbf{r}\}| \leq \min(\beta, |G| - 1)$ *and*
$\forall i \in G, |\{o \in T \text{ such that } (o, i) \notin \mathbf{r}\}| \leq \min(\alpha, |T| - 1)$.

Example 4. *Given $\mathbf{r}_1$ and $\alpha = \beta = 1$, the two bi-sets $(\{o_1, o_2, o_3\}, \{i_1, i_2\})$ and $(\{o_1, o_2, o_4\}, \{i_1, i_2\})$ satisfy the $\alpha\beta$-constraint. However, o_3 and o_4 have the same values on i_1 and i_2. It turns out that these objects can not be added simultaneously on $(\{o_1, o_2\}, \{i_1, i_2\})$ in order to satisfy $\mathcal{C}_{\alpha\beta}$.*

To ensure consistency and avoid this problem, we decided either to add all identical properties (w.r.t. the set of objects) or all identical objects (w.r.t. the set of properties) in the bi-set when $\mathcal{C}_{\alpha\beta}$ is satisfied, or to exclude all of them when it is not the case. As for concepts, α/β concepts can differ from each other either on the object component or on the property component. This is formalized by the use of the consistency constraint denoted $\mathcal{C}_{cons}$.

Definition 8 (Consistency Constraint). *A bi-set (T, G) satisfies $\mathcal{C}_{cons}$ iff*

- $\forall i \in G,\ \nexists j \in \mathcal{P} \setminus G$ *such that* $\psi(i) \cap T = \psi(j) \cap T$
- $\forall o \in T,\ \nexists w \in \mathcal{O} \setminus T$ *such that* $\phi(o) \cap G = \phi(w) \cap G$

On our way to the extraction of bi-sets with few 0 values, it is interesting to reformulate the definition of formal concepts.

Definition 9 (Maximality Constraint). *A bi-set (T, G) is maximal w.r.t. a constraint $\mathcal{C}$ and is said to satisfy $\mathcal{C}_{max|\mathcal{C}}(T, G)$ iff $\nexists (T', G')$ such that $\mathcal{C}(T', G') \wedge (T, G) \prec (T', G')$.*

Definition 10 (New Definition of Formal Concepts). *A bi-set (T, G) is a formal concept iff*

- *(T, G) satisfies $\mathcal{C}_{1R}$*
- *(T, G) is maximal w.r.t. $\mathcal{C}_{1R}$, i.e., (T, G) satisfies $\mathcal{C}_{max|\mathcal{C}_{1R}}$.*

Notice that by construction, a concept satisfies the constraint $\mathcal{C}_{cons}$. Let us now define α/β concepts.

Definition 11 (α/β Concept). *A bi-set (T, G) is an α/β concept iff*

- *(T, G) satisfies $\mathcal{C}_{\alpha\beta}$*
- *(T, G) satisfies $\mathcal{C}_{cons}$*
- *(T, G) is maximal w.r.t. $\mathcal{C}_{\alpha\beta} \wedge \mathcal{C}_{cons}$, i.e., (T, G) satisfies $\mathcal{C}_{max|\mathcal{C}_{\alpha\beta} \wedge \mathcal{C}_{cons}}$.*

Let us notice that, looking for an α/β concept (T, G), it makes sense that $|T| \gg \alpha$ and $|G| \gg \beta$. The $\alpha\beta$-constraint is an extension of the 1-rectangle constraint for bi-sets with 0 values. Then, α/β concepts appear to be a simple extension of concepts by changing the 1-rectangle constraint into the $\alpha\beta$-constraint in conjunction with the $\mathcal{C}_{cons}$ constraint. This is one of the important results of this work.

Example 5. *$(\{o_1, o_2, o_3\}, \{i_1, i_2, i_3\})$ is an α/β concept in $\mathbf{r}_1$. $(\{o_1, o_2\}, \emptyset)$ and $(\{o_3, o_4, o_5\}, \{i_1, i_2\})$ are not α/β concepts because they do not satisfy respectively $\mathcal{C}_{max|\mathcal{C}_{cons} \wedge \mathcal{C}_{\alpha\beta}}$ and $\mathcal{C}_{\alpha\beta}$ constraints.*

3 Mining α/β Concepts

The computation of every α/β concept from a given data set $\mathbf{r}$ is done in two steps. First, we compute all the concepts, i.e., a collection denoted $\mathcal{K}$. Then we search the maximal (w.r.t. a specialization relation on bi-sets) unions of concepts which satisfy the $\alpha\beta$-constraint $\mathcal{C}_{\alpha\beta}$.

Definition 12 (Union of Bi-sets). *Let $B_1 = (T_1, G_1)$ and $B_2 = (T_2, G_2)$ be two bi-sets from $\mathcal{L_O} \times \mathcal{L_P}$. The union of B_1 and B_2 is $B_1 \sqcup B_2 = (T_1 \cup T_2, G_1 \cup G_2)$. It can be applied on concepts that are special cases of bi-sets. By construction, unions of concepts are not concepts.*

Theorem 1. *Let $U = \{\bigsqcup_{i \in \mathcal{K}'} i \mid \mathcal{C}_{\alpha\beta}$ and $\mathcal{K}' \subseteq \mathcal{K}\}$ where $\mathcal{K}$ is the collection of concepts, the collection of α/β concepts is equal to*

$$\mathcal{K}_{\alpha\beta} = \{s \in U \mid \nexists s\prime \in U\ s \preceq s\prime\}$$

Proof. We show that the collection of bi-sets which satisfy $\mathcal{C}_{cons}$ ($\mathcal{K}_{cons}$) is equal to the collection of the unions of concepts ($\mathcal{K}_{\sqcup}$). In other terms, the use of unions enforce the $\mathcal{C}_{cons}$ constraint.

- $\mathcal{K}_{\sqcup} \subseteq \mathcal{K}_{cons}$
  Let (X, Y) be an element of $\mathcal{K}_{\sqcup}$. Let us assume that $\neg\mathcal{C}_{cons}(X, Y)$. We consider $j \in \mathcal{P} \setminus Y$ such that $\exists i \in Y$, $\psi(i) \cap X = \psi(j) \cap X$. It exists at least one concept $(L, C) \in \mathcal{K}$ such that $(L, C) \preceq (X, Y)$ and $i \in C$ ((X, Y) is a union of concepts). However, $\forall \ell \in L$, $(\ell, i) \in \mathbf{r}$ and $L \subseteq \psi(j)$, thus $(\ell, j) \in \mathbf{r}$. Consequently, as (L, C) is a concept, $j \in C \subseteq Y$. We have a contradiction and $\mathcal{C}_{cons}$ is satisfied.
  Reciprocally, we consider $w \in \mathcal{O} \setminus X$ such that $\exists v \in X$, $\phi(v) \cap Y = \phi(w) \cap Y$. It exists at least one concept $(L, C) \in \mathcal{K}$ such that $(L, C) \preceq (X, Y)$ and $v \in L$ ((X, Y) is a union of concepts). However, $\forall c \in C$, $(v, c) \in \mathbf{r}$ *and* $C \subseteq \phi(w)$, thus $(w, c) \in \mathbf{r}$. Consequently, as (L, C) is a concept, $w \in L \subseteq X$. We have a contradiction and thus $\mathcal{C}_{cons}$ is satisfied.
- $\mathcal{K}_{cons} \subseteq \mathcal{K}_{\sqcup}$
  Let (X, Y) be a bi-set which satisfy $\mathcal{C}_{cons}$. $\forall i \in Y$, $\psi(i) \cap X \neq \emptyset$ and $\nexists j \in \mathcal{P} \setminus Y$ such that $\psi(i) \cap X = \psi(j) \cap X$ consequently $\phi(\psi(i) \cap X) \subseteq Y$. As $\psi(i) \cap X \subseteq \psi(i)$ and ϕ is a decreasing operator, $\phi(\psi(i)) \subseteq \phi(\psi(i) \cap X)$ consequently $\phi(\psi(i)) \subseteq Y$.
  On the other side, $\psi(i) \cap X \neq \emptyset$. Let $v \in \psi(i) \cap X$. It does not exist $w \in \mathcal{O} \setminus \psi(i) \cap X$ such that $\phi(v) \cap Y = \phi(w) \cap Y$ consequently $\psi(\phi(v) \cap Y) \subseteq X$. As $\phi(v) \cap Y \subseteq \phi(v)$ and ψ is a decreasing operator, $\psi(\phi(v)) \subseteq \psi(\phi(v) \cap Y)$ consequently $\psi(\phi(v)) \subseteq X$.
  We can conclude that for each $(v, i) \in (X, Y)$ and $(v, i) \in \mathbf{r}$, it exists a concept $(\psi(\phi(v)), \phi(\psi(i))$ included in (X, Y). (X, Y) is the union of these concepts and thus belongs to $\mathcal{C}_{\sqcup}$.

It means that we can compute α/β concepts by generating the unions of concepts which satisfy $\mathcal{C}_{\alpha\beta}$ and $\mathcal{C}_{max|\mathcal{C}_{\alpha\beta}}$.

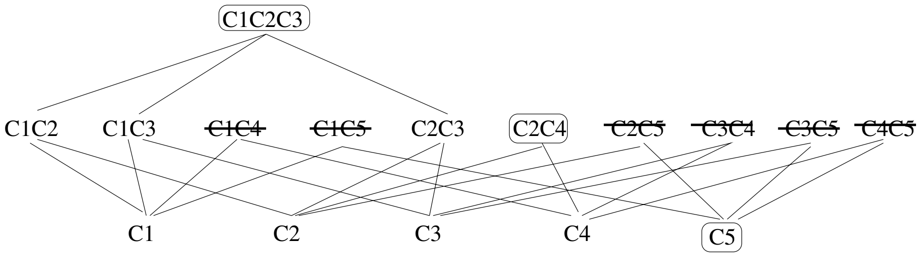

Fig. 2. Search space of α/β concepts ($\alpha = \beta = 1$) in context $\mathbf{r}_1$

Property 1. $\mathcal{C}_{\alpha\beta}$ is anti-monotonic w.r.t. our specialization relation on bi-sets.

Consequently, when considering candidate unions of concepts, we can use the anti-monotonicity of $\mathcal{C}_{\alpha\beta}$ to prune the search space. It is also possible to push $\mathcal{C}_{max|\mathcal{C}_{\alpha\beta}}$ to prune the search space. This can be done by adapting known efficient algorithms which compute maximal frequent sets (see, e.g., [10] for a recent survey), where sets are sets of concepts and the anti-monotonic minimal frequency constraint is replaced by the $\mathcal{C}_{\alpha\beta}$ constraint.

Given $\mathcal{K}$ the collection of formal concepts and two parameters α and β, we compute the following collection of sets of concepts $\{\varphi \in 2^{\mathcal{K}} \mid \mathcal{C}_{\alpha\beta}(\varphi) \wedge \mathcal{C}_{max|\mathcal{C}_{\alpha\beta}}(\varphi)$ is satisfied$\}$.

The concepts in $\mathbf{r}_1$ are:

$$\begin{array}{rl} c_1 = (\{o_1\}, \{i_1, i_2, i_3\}) & c_2 = (\{o_1, o_2\}, \{i_1, i_2\}) \\ c_3 = (\{o_1, o_3\}, \{i_1, i_3\}) & c_4 = (\{o_1, o_2, o_5\}, \{i_2\}) \\ c_5 = (\{o_1, o_2, o_3, o_4\}, \{i_1\}) & \end{array}$$

We consider the search for α/β concepts in $\mathbf{r}_1$ when $\alpha = 1$ and $\beta = 1$.

Figure 2 illustrates how the collection of 1/1 concepts are extracted from $\mathbf{r}_1$: it provides $\{\{c_1, c_2, c_3\}, \{c_2, c_4\}, \{c_5\}\}$. The circled elements form the solution space. Stripped elements do not satisfy $\mathcal{C}_{\alpha\beta}$. Their supersets are not generated. The three α/β concepts are here $c_1 \cup c_2 \cup c_3$, $c_2 \cup c_4$, and c_5. They correspond respectively to the following bi-sets: $(\{o_1, o_2, o_3\}, \{i_1, i_2, i_3\})$, $(\{o_1, o_2, o_5\}, \{i_1, i_2\})$ and $(\{o_1, o_2, o_3, o_4\}, \{i_1\})$.

4 Experimentation

4.1 Synthetic Data

To show the relevancy of α/β concept mining in noisy data, we first designed a synthetic data set. Our goal was to show that α/β concept mining enables to discover concepts that have been introduced before the introduction of some noise. Therefore, we have built a boolean data set made of 20 non-overlapping concepts containing each 5 items and 5 objects. Secondly, we introduced a uniform random noise by modifying with the same probability (5% in Figure 3 top

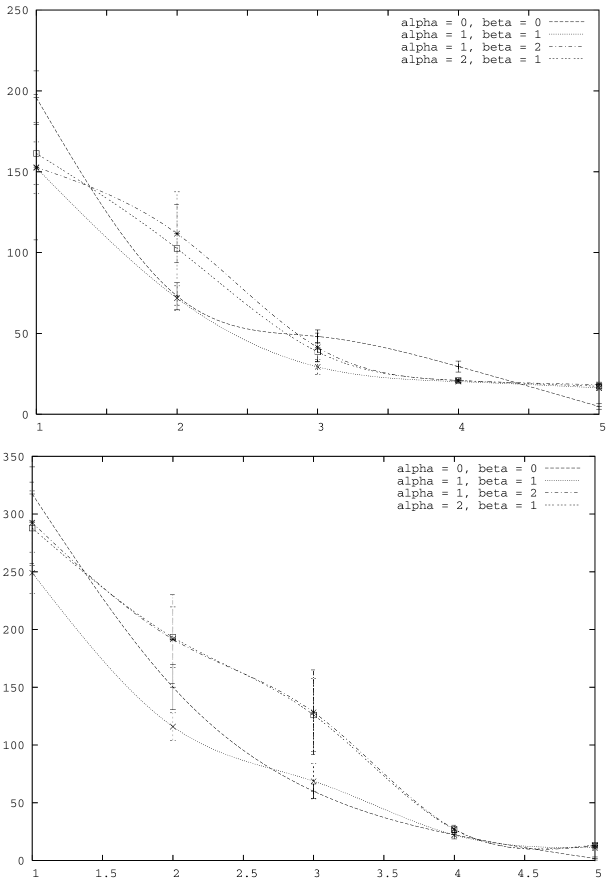

Fig. 3. Number of α/β concepts with respect to their size (both dimensions greater or equal than the X-coordinate value) with 5% (top) and 10% (bottom) of noise

and 10% in Figure 3 bottom) some of the boolean values (i.e., transforming some true values into false values and vice versa). We produced 10 data sets (with 100 lines and 100 columns) for each noise probability. When considering first concept mining phase, we extracted between 169 and 225 concepts (resp. between 289 and 373 concepts) in the 5% noise data sets (resp. in the 10% noise data sets). Figure 3 provides the average and standard deviation of the number of α/β concepts (Y-coordinate) w.r.t. their minimal number of objects and properties (X-coordinate). Each curve stands for a different value of α and β between 0 and 2. For example, on Figure 3 bottom, we have 126 α/β concepts in average with at least 3 objects and 3 items when $\alpha = 2$ and $\beta = 1$.

On the data sets with 5% noise, we have in average 196 concepts (see the curve with $\alpha = \beta = 0$) among which 48 have at least 3 properties and objects and 5 of them have at least 5 properties and objects. With 10% of noise, we got 317 concepts in average among which 60 have at least 3 properties and objects and

2 of them have at least 5 properties and objects. In this extracted collection of concepts, it is difficult to find the 20 original concepts that were occurring before noise introduction. When α and β are not null, the collection of extracted α/β concepts is roughly speaking the 20 original concepts. For example, considering $\alpha = \beta = 1$, we got 20.2 (resp. 22.1) α/β concepts of size greater than 4 in the 5% (resp. the 10%) noise data set. Even when the percentage of noise increases, the collection of α/β concepts has "captured" the embedded concepts. Nevertheless, the number of α/β concepts can increase with α or β. A lot of α/β concepts with a number of objects close to α and a number of properties close to β leads to the computation of many unions. However, when several unions have been performed, it is more and more difficult to merge concepts. α/β concepts whose the minimal number of lines and columns is large w.r.t. α and β are dense in terms of true values and considered relevant. In other terms, it is interesting not to consider small α/β concepts (w.r.t. α and β) and thus eliminate lots of meaningless α/β concepts.

4.2 Post-processing an Incomplete Collection of Concepts on Real Gene Expression Data

In many real data sets, it is not possible to extract the whole collection of concepts. In these cases, additional constraints can be pushed deeply into the concept extraction algorithms like, for instance, enforcing a minimal size for both set components when using our D-Miner algorithm [3, 4]. We could also limit the search to the so-called frequent concepts which use such a constraint on only one set component (see, e.g., [18, 15]).

Even in the case where we can not have the whole collection of concepts $\mathcal{K}$, we can still extract α/β concepts from a subset of $\mathcal{K}$. Doing so, we compute more relevant patterns as a post-processing of some concepts.

A concrete application concerns the analysis of gene expression profiles in Type 2 diabetes disease. As we already pointed out, molecular biologists are interested in discovering putative transcription modules, i.e. sets of genes that are co-regulated and the associated sets of situations where this co-regulation occurs. In the following experiment, situations corresponds to transcription factors, i.e. biological objects which are known to activate or repress the genes. We derived a boolean data set from the data in [17]. It contains 350 genes (in rows) which are in relation with some transcription factors (150 columns) known to regulate (activate or repress) them. This data set is dense since 17% of the values are true values.

We are interested in large α/β concepts that associate many genes to many transcription factors. We were not able to extract the collection of α/β concepts from the whole collection of concepts (more than 5 millions). We decided to look at the merging of large concepts containing at least 25 genes and 10 transcription factors. Using D-Miner, we extracted 1 699 concepts satisfying these size constraints. Then we computed the collections of α/β concepts with small α and β values. Table 1 provides the number of α/β concepts (for 4 values of $\alpha\beta$) per number of merged concepts.

Table 1. Number of α/β concepts produces by the union of n concepts

n	$\alpha = \beta = 1$	$\alpha = \beta = 2$	$\alpha = \beta = 3$	$\alpha = \beta = 4$
1	1450	1217	927	639
2	54	49	61	95
3	31	57	75	73
4	8	40	50	64
5	2	8	25	58
6	1	3	11	29
7	0	0	6	11
8	0	0	1	12
9	0	0	0	2
10	0	0	1	6
11	0	0	0	0
12	0	0	0	3
13	0	0	0	1
14	0	0	1	0
15	0	0	0	1
Total	1546	1374	1158	994

Table 2. α/β concept (36×12) resulting from the union of 15 concepts with $\alpha = \beta = 4$ (number of false values for each transcription factor of the α/β concept)

Number of false values
0
0
0
1
3
2
0
3
0
3
4
0

Interestingly, even though we merged only large concepts with small α and β values, large α/β concepts have appeared. For example, at most 6 concepts are merged when $\alpha = \beta = 1$ whereas 15 concepts are merged when $\alpha = \beta = 4$. In this data set, we have large bi-sets with few 0 values. Typically, the α/β concept ($\alpha = \beta = 4$) resulting from the merge of 15 concepts is made of 36 genes and 12 transcription factors and contains only 3.7% of false values (see Table 2 where each line stands for a transcription factor and the value is the number of false values in the α/β concept).

The 12 transcription factors of this α/β concept have been checked as really similar with respect to the genes which are associated. It seems useful for biologists to consider such α/β concepts with very few exceptions instead of post-processing by themselves huge collections of concepts.

5 Conclusion

We have considered the challenging problem of computing formal concepts with exceptions from transactional data sets. This is extremely important in many application domains where strongly associated sets of objets and properties can provide interesting patterns. Closed sets associated via the Galois connection are indeed strongly associated but we miss interesting associations when the data is intrinsically noisy, for instance because of measurement errors or some crispy discretization procedures. The same reasoning has lead few years ago to the computation of almost-closure [5] when looking for condensed representations of frequent itemsets. The difficulty here has been to design a complete method for computing the so-called α/β concepts. Our formalization in terms of union of concepts that satisfy $\mathcal{C}_{\alpha\beta}$ and $\mathcal{C}_{max|\mathcal{C}_{\alpha\beta}}$ is complete. We experimentally validated the added-value of the approach on both synthetic data and a real application in molecular biology. Further experiments are needed for a better understanding of the difference between collections of concepts and collections of α/β concepts.

Acknowledgement. Jérémy Besson is funded by INRA. The authors thank Sophie Rome and Christophe Rigotti for stimulating discussions and constructive feedback during the preparation of this paper.

References

1. C. Becquet, S. Blachon, B. Jeudy, J.-F. Boulicaut, and O. Gandrillon. Strong association rule mining for large gene expression data analysis: a case study on human SAGE data. *Genome Biology*, 12, 2002. See http://genomebiology.com/2002/3/12/research/0067.
2. A. Berry, J.-P. Bordat, and A. Sigayret. Concepts can not afford to stammer. In *Proceedings JIM'03*, pages 25–35, Metz, France, September 2003.
3. J. Besson, C. Robardet, and J.-F. Boulicaut. Constraint-based mining of formal concepts in transactional data. In *Proceedings PaKDD'04*, volume 3056 of *LNCS*, pages 615–624, Sydney, Australia, May 2004. Springer-Verlag.
4. J. Besson, C. Robardet, J.-F. Boulicaut, and S. Rome. Constraint-based bi-set mining for biologically relevant pattern discovery in microarray data. *Intelligent Data Analysis journal*, 9, 2004. In Press.
5. J.-F. Boulicaut and A. Bykowski. Frequent closures as a concise representation for binary data mining. In *Proceedings PaKDD'00*, volume 1805 of *LNAI*, pages 62–73, Kyoto, JP, Apr. 2000. Springer-Verlag.
6. J.-F. Boulicaut, A. Bykowski, and C. Rigotti. Free-sets: a condensed representation of boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery journal*, 7(1):5–22, 2003.
7. A. Bykowski and C. Rigotti. DBC: a condensed representation of frequent patterns for efficient mining. *Information Systems Journal*, 28(8):949–977, 2003.
8. B. Ganter. Two basic algorithms in concept analysis. Technical report, Technisch Hochschule Darmstadt, Preprint 831, 1984.
9. F. Geerts, B. Goethals, and T. Mielikäinen. Tiling databases. In *Proceedings DS'04*, volume 3245 of *LNCS*, Padova, Italy, Oct. 2004. Springer-Verlag. To appear.

10. B. Goethals and M. J. Zaki, editors. *FIMI '03, Frequent Itemset Mining Implementations, Proceedings of the ICDM 2003 Workshop on Frequent Itemset Mining Implementations, 19 December 2003, Melbourne, Florida, USA*, volume 90 of *CEUR Workshop Proceedings*, 2003.
11. L. Nourine and O. Raynaud. A fast algortihm for building lattices. *Information Processing Letters*, 71:190–204, 1999.
12. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24(1):25–46, Jan. 1999.
13. J. Pei, J. Han, and R. Mao. CLOSET an efficient algorithm for mining frequent closed itemsets. In *Proceedings ACM SIGMOD Workshop DMKD'00*, 2000.
14. R. Pensa, C. Leschi, J. Besson, and J.-F. Boulicaut. Assessment of discretization techniques for relevant pattern discovery from gene expression data. In *Proceedings BIOKDD'04 co-located with ACM SIGKDD'04*, Seattle, USA, August 2004. In Press.
15. F. Rioult, J.-F. Boulicaut, B. Crémilleux, and J. Besson. Using transposition for pattern discovery from microarray data. In *Proceedings ACM SIGMOD Workshop DMKD'03*, pages 73–79, San Diego, USA, June 2003.
16. F. Rioult, C. Robardet, S. Blachon, B. Crémilleux, O. Gandrillon, and J.-F. Boulicaut. Mining concepts from large SAGE gene expression matrices. In *Proceedings KDID'03 co-located with ECML-PKDD'03 ISBN:953-6690-34-9*, pages 107–118, Cavtat-Dubrovnik, Croatia, September 22 2003.
17. S. Rome, K. Clément, R. Rabasa-Lhoret, E. Loizon, C. Poitou, G. S. Barsh, J.-P. Riou, M. Laville, and H. Vidal. Microarray profiling of human skeletal muscle reveals that insulin regulates 800 genes during an hyperinsulinemic clamp. *Journal of Biological Chemistry*, March 2003. In Press.
18. G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal. Computing iceberg concept lattices with titanic. *Data and Knowledge Engineering*, 42:189–222, 2002.
19. J. Wang, J. Han, and J. Pei. CLOSET+: searching for the best strategies for mining frequent closed itemsets. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003.
20. R. Wille. Restructuring lattice theory: an approach based on hierarchies of concepts. In I. Rival, editor, *Ordered sets*, pages 445–470. Reidel, 1982.
21. M. J. Zaki and C.-J. Hsiao. CHARM: An efficient algorithm for closed itemset mining. In *Proceedings SIAM DM'02*, Arlington, USA, April 2002.