From Local Pattern Mining to Relevant Bi-cluster Characterization

Ruggero G. Pensa and Jean-François Boulicaut

INSA Lyon, LIRIS CNRS,
UMR 5205 F-69621, Villeurbanne cedex, France
{Ruggero.Pensa, Jean-Francois.Boulicaut}@insa-lyon.fr

Abstract. Clustering or bi-clustering techniques have been proved quite useful in many application domains. A weakness of these techniques remains the poor support for grouping characterization. We consider eventually large Boolean data sets which record properties of objects and we assume that a bi-partition is available. We introduce a generic cluster characterization technique which is based on collections of bi-sets (i.e., sets of objects associated to sets of properties) which satisfy some user-defined constraints, and a measure of the accuracy of a given bi-set as a bi-cluster characterization pattern. The method is illustrated on both formal concepts (i.e., "maximal rectangles of true values") and the new type of δ-bi-sets (i.e., "rectangles of true values with a bounded number of exceptions per column"). The added-value is illustrated on benchmark data and two real data sets which are intrinsically noisy: a medical data about meningitis and Plasmodium falciparum gene expression data.

1 Introduction

Clustering has been proved extremely useful for exploratory data analysis. Its main goal is to identify a partition of objects and/or properties such that an objective function which specifies its quality is optimized (e.g., maximizing intra-cluster similarity and inter-cluster dissimilarity). Looking for optimal solutions is intractable such that heuristic local search optimizations are performed [1]. Many efficient algorithms can provide good partitions but suffer from the lack of explicit cluster characterization. For example, considering gene expression data analysis, clustering is used to look for sets of co-expressed genes and/or sets of biological situations or experiments which seem to trigger this co-expression (see, e.g., [2]). In this context, an explicit characterization would be a symbolic statement which "explains" why genes and/or situations are within the same groups. Once such characterizations are available, it supports the understanding of gene regulation mechanisms. Our running example $\mathbf{r}$ (see Table 1) concerns a toy Boolean data set. For instance, it encodes gene expression properties (e.g., over-expression) in various biological situations and, genes denoted by p_1, p_3, p_4 are considered over-expressed in situation o_1.

The crucial need for characterization has motivated the research on conceptual clustering [3]. Among others, it has been studied in the context of co-clustering or bi-clustering [4,5,6,7], including for the special case of categorial

A.F. Famili et al. (Eds.): IDA 2005, LNCS 3646, pp. 293–304, 2005.

Table 1. A Boolean context **r**

	p_1	p_2	p_3	p_4	p_5
o_1	1	0	1	1	0
o_2	0	1	0	0	1
o_3	1	0	1	1	0
o_4	0	0	1	1	0
o_5	1	1	0	0	1
o_6	0	1	0	0	1
o_7	0	0	0	0	1

or Boolean data. The goal is to identify bi-clusters or bi-partitions in the data, i.e., a mapping between a partition of situations (more generally objects) and a partition of gene expression properties (more generally, Boolean properties of objects). For instance, an algorithm like Cocluster [6] can compute the interesting bi-partition $\{\{\{o_1, o_3, o_4\}, \{p_1, p_3, p_4\}\}, \{\{o_2, o_5, o_6, o_7\}, \{p_2, p_5\}\}\}$ from **r**. The first bi-cluster indicates that the characterization of objects from $\{o_1, o_3, o_4\}$ is that they almost always share properties from $\{p_1, p_3, p_4\}$. Also, properties in $\{p_2, p_5\}$ are characteristics for objects in $\{o_2, o_5, o_6, o_7\}$. Unfortunately, this first step towards characterization is not sufficient to support the needed interactivity with the end-users who have to interpret the resulting (bi-)partitions. Our thesis is that it is useful to look for bi-sets, i.e., sets of objects associated to sets of properties, that exhibit strong and characteristic relations between bi-cluster elements. For instance, once a bi-partition of a Boolean gene expression data set has been found, one can be interested in studying all the interactions between genes involved in a "cancer" bi-cluster, and these interactions might imply genes which are involved in "non cancerous" processes as well.

Given a bi-partition on a Boolean data set, our goal is to provide characterizing patterns for each bi-cluster and our contribution is twofold. First, we introduce an original and generic cluster characterization technique which is based on constraint-based bi-set mining, i.e., mining bi-sets whose set components satisfy some constraints, and a measure of the accuracy of a given extracted bi-set as a characterization pattern for a given bi-cluster (see Section 2). We also discuss the opportunity to shift from the characterization by bi-sets towards a characterization based on association rules. The method is then illustrated on two kinds of bi-sets, the well-known formal concepts (i.e., associated closed sets [8] or, intuitively, "maximal rectangle of true values") and a new class, the so-called δ-bi-sets. This later pattern type is new and it is based on a previous work about approximate condensed representations for frequent patterns [9]. Intuitively, a δ-bi-set is a "rectangle of true values with a bounded number of exceptions per column" (see Section 3). We illustrate the added-value of our characterizing method not only on a benchmark data set but also on two real-life data sets. The obtained characterizations are consistent with the available knowledge (see Section 4).

2 Bi-cluster Characterization Using Bi-sets

Let us consider a set of objects $\mathcal{O} = \{o_1, \ldots, o_m\}$ and a set of Boolean properties $\mathcal{P} = \{p_1, \ldots, p_n\}$. The Boolean context to be mined is $\mathbf{r} \subseteq \mathcal{O} \times \mathcal{P}$, where $r_{ij} = 1$ if the property p_j is true for object o_i. Formally, a bi-set is an element of $2^{\mathcal{O}} \times 2^{\mathcal{P}}$. We assume that a bi-clustering algorithm, e.g., [6], provides a mapping between k clusters of objects (say $\{C_1^o \ldots C_k^o\}$) and k clusters of properties (say $\{C_1^p \ldots C_k^p\}$). A first characterization comes from this mapping.

Our goal is to support each bi-cluster interpretation by collections of bi-sets which are locally pointing out interesting associations between groups of objects and groups of properties. Therefore, we assume that a collection of N bi-sets $\mathcal{C} = c_1, \ldots, c_N$ has been extracted from the data. First, we associate each of them to one the k bi-clusters to obtain a collection of k groups of bi-sets $\{C_1, \ldots, C_k\}$, where $C_i \subseteq \mathcal{C}$. Each bi-set $\in C_i$ characterizes the bi-cluster (C_i^o, C_i^p) with some degree of accuracy.

Let us first define the signature in $\mathbf{r}$ of each bi-cluster (C^o, C^p) denoted $\mu(C^o, C^p) = (\boldsymbol{\tau}, \boldsymbol{\gamma})$ where $\tau = \{o_i \in C^o\}$ and $\gamma = \{p_i \in C^p\}$. We can now define a similarity measure between a bi-set $c = (T, G)$ and a bi-cluster signature:

$$sim(c, \mu(C^o, C^p)) = \frac{|(T, G) \cap (\boldsymbol{\tau}, \boldsymbol{\gamma})|}{|(T, G) \cup (\boldsymbol{\tau}, \boldsymbol{\gamma})|} = \frac{|T \cap \boldsymbol{\tau}| \cdot |G \cap \boldsymbol{\gamma}|}{|T| \cdot |G| + |\boldsymbol{\tau}| \cdot |\boldsymbol{\gamma}| - |T \cap \boldsymbol{\tau}| \cdot |G \cap \boldsymbol{\gamma}|}$$

Intuitively, bi-sets (T, G) and $(\boldsymbol{\tau}, \boldsymbol{\gamma})$ denote rectangles in the matrix (modulo permutations over the lines and the columns) and we measure the area of the intersection of the two rectangles normalized by the area of their union.

Each bi-set c which is a candidate characterization pattern can now be assigned to the bi-cluster (C^o, C^p) for which $sim(c, \mu(C^o, C^p))$ is maximal. Doing so, we get k groups of potentially characterizing bi-sets. Finally, we can use an accuracy measure to select the most relevant ones. For that purpose, we propose to measure the exception ratios for the two set components of the bi-sets.

Given a bi-set (T, G) and a bi-cluster (C^o, C^p), it can be computed as follows:

$$\epsilon_o = \frac{|\{o_i \in T |\ o_i \notin C^o\}|}{|T|}, \quad \epsilon_p = \frac{|\{p_i \in G |\ p_i \notin C^p\}|}{|G|}$$

It is then possible to consider thresholds to select only the bi-sets that have little exception ratios, i.e., $\epsilon_o < \varepsilon_o$ and $\epsilon_p < \varepsilon_p$ where $\varepsilon_o, \varepsilon_p \in [0, 1]$. There are several possible interpretations for these measures. If we are interested in characterizing a cluster of objects (resp. properties), we can look for all the sets of properties (resp. objects) for which the ϵ_o (resp. ϵ_p) values of the related bi-sets are less than a threshold ε_o (resp. ε_p). Alternatively, we can consider the whole bi-cluster and characterize it with all the bi-sets for which the two exception ratios ϵ_o and ϵ_p are less than two threshold ε_o and ε_p.

3 Looking for Candidate Characterizing Bi-sets

We now discuss the type of bi-sets which will be post-processed for bi-cluster characterization. It is clear that bi-clusters are, by construction, interesting char-

acterizing bi-sets but they only support a global interpretation. We are interested in strong associations between sets of objects and sets of properties that can locally explain the global behavior. Clearly, formal concepts can be used [8].

Definition 1 (formal concept). *If $T \subseteq \mathcal{O}$ and $G \subseteq \mathcal{P}$, assume $\phi(T, \mathbf{r}) = \{g \in \mathcal{P} \mid \forall t \in T, (t, g) \in \mathbf{r}\}$ and $\psi(G, \mathbf{r}) = \{t \in \mathcal{O} \mid \forall g \in G, (t, g) \in \mathbf{r}\}$. A bi-set (T, G) is a formal concept in $\mathbf{r}$ when $T = \psi(G, \mathbf{r})$ and $G = \phi(T, \mathbf{r})$. By construction, G and T are closed sets, i.e., $G = \phi \circ \psi(G, \mathbf{r})$ and $T = \psi \circ \phi(T, \mathbf{r})$. Intuitively, (T, G) is a maximal rectangle of true values.*

$(\{o_1, o_3\}, \{p_1, p_3, p_4\})$, $(\{o_1, o_3, o_4\}, \{p_3, p_4\})$, and $(\{o_5, o_6\}, \{p_2, p_5\})$ are examples of formal concepts among the 8 ones which hold in $\mathbf{r}$ (see Table 1). Efficient algorithms have been developed to extract complete collections of formal concepts which satisfy also user-defined constraints, e.g., [10,11]. A fundamental problem with formal concepts is that the Galois connection (ϕ, ψ) is, in some sense, a too strong one: we have to capture every maximal set of objects and its maximal set of associated properties. As a result, the number of formal concepts even in small matrices can be huge. A solution is to look for "dense" rectangles in the matrix, i.e., bi-sets with mainly true values but also a bounded (and small) number of false values or exceptions. Well-defined collections of dense bi-sets can be obtained by merging formal concepts [12], i.e., a post-processing over collections of formal concepts. This turns to be intractable when the number of formal concepts is too large. We propose a new type of bi-set which can be efficiently extracted, including in noisy data in which it is common to have several millions of formal concepts.

3.1 Mining δ-Bi-sets

We want to compute efficiently smaller collections of bi-sets which still capture strong associations. We recall some definitions about the association rule mining task [13] since it is used for both the definition of the δ-bi-set pattern type and for bi-cluster characterization.

Definition 2 (association rule, frequency, confidence). *An association rule R in $\mathbf{r}$ is an expression of the form $X \Rightarrow Y$, where $X, Y \subseteq \mathcal{P}$, $Y \neq \emptyset$ and $X \cap Y = \emptyset$. Its absolute frequency is $|\psi(X \cup Y, \mathbf{r})|$ and its confidence is $|\psi(X \cup Y, \mathbf{r})|/|\psi(X, \mathbf{r})|$.*

In an association rule $X \Rightarrow Y$ with high confidence, the properties in Y are almost always true for an object when the properties in X are true. Intuitively, $X \cup Y$ associated to $\psi(X, \mathbf{r})$ is then a dense bi-set: it contains a few false values. We now consider our technique for computing association rules with high confidence, the so-called δ-strong rules [14,9].

Definition 3 (δ-strong rule). *Given an integer δ, a δ-strong rule in $\mathbf{r}$ is an association rule $X \Rightarrow Y$ $(X, Y \subset \mathcal{P})$ s.t. $|\psi(X, \mathbf{r})| - |\psi(X \cup Y, \mathbf{r})| \leq \delta$, i.e., the rule is violated in no more than δ objects.*

Interesting collections of δ-strong rules with minimal left-hand side can be computed efficiently from the so-called δ-free-sets [14,9,15] and their δ-closures.

Definition 4 (δ-free set, δ-closure). *Let δ be an integer and $X \subset \mathcal{P}$, X is a δ-free-set in $\mathbf{r}$ iff there is no δ-strong rule which holds between two of its own and proper subsets. The δ-closure of X in $\mathbf{r}$, $h_\delta(X, \mathbf{r})$, is the maximal superset Y of X s.t. $\forall p \in Y \setminus X$, $|\psi(X \cup \{p\})| \geq |\psi(X, \mathbf{r})| - \delta$. In other terms, the frequency of the δ-closure of X in $\mathbf{r}$ is almost the same than the frequency of X when $\delta << |\mathcal{O}|$. Moreover, $\forall p \in h_\delta(X) \setminus X$, $X \Rightarrow p$ is a δ-strong rule.*

For example, in Table 1, the 1-free itemsets are $\{p_1\}$, $\{p_2\}$, $\{p_3\}$, $\{p_4\}$, $\{p_5\}$, $\{p_1, p_2\}$, and $\{p_1, p_5\}$. An example of 1-closure for $\{p_1\}$ is $\{p_3, p_4\}$. The association rules $\{p_1\} \Rightarrow \{p_3\}$ and $\{p_1\} \Rightarrow \{p_4\}$ have only one exception.

δ-freeness is an anti-monotonic property such that it is possible to compute δ-free sets (eventually combined with a minimal frequency constraint) in very large data sets. Notice that $h_0 \equiv \phi \circ \psi$, i.e., the classical closure operator. Looking for a 0-free-set, say X, and its 0-closure, say Y, provides the closed set $X \cup Y$ and thus the formal concept $(\psi(X \cup Y, \mathbf{r}), X \cup Y)$.

Definition 5 (δ-bi-set). *A δ-bi-set (T, G) in $\mathbf{r}$ is built on each δ-free-set $X \subset \mathcal{P}$ with $T = \psi(X, \mathbf{r})$ and $G = h_\delta(X, \mathbf{r})$.*

In Table 1, the 1-bi-sets derived from the 1-free-sets $\{p_3\}$ and $\{p_5\}$ are $(\{o_1, o_3, o_4\}, \{p_1, p_3, p_4\})$ and $(\{o_2, o_5, o_6, o_7\}, \{p_2, p_5\})$. When $\delta << |T|$, δ-bi-sets are dense bi-sets with a small number of exceptions per column. In order to experiment, we implemented a straightforward extension of AcMiner [9] which provides the supporting set for each extracted δ-free-set.

3.2 Concepts vs. δ-Bi-sets

To study the relevancy of δ-bi-sets w.r.t. formal concepts, we have considered the addition of noise to a synthetical data set. Hereafter, $\mathbf{r}$ denotes a reference data set from which we generate noisy data sets by adding a given quantity of uniform random noise. Then, we compare the collection of formal concepts which are "built-in" within $\mathbf{r}$ with various collections of formal concepts and δ-bi-sets extracted from the noised matrices. To measure the relevancy of each extracted collection w.r.t the reference one, we look for subsets of the reference collection in each of them. Since both set components of each formal concept can be changed when adding noise, we identify those having the largest area in common with the reference ones, and we compute the σ measure which takes into account the common area:

$$\sigma(\mathcal{C}_r, \mathcal{C}_a) = \frac{1}{N_r} \sum_{i=1}^{N_r} max_j \left(\frac{|(T_i, G_i)_r \cap (T_j, G_j)_a|}{|(T_i, G_i)_r \cup (T_j, G_j)_a|} \right)$$

where $\mathcal{C}_r$ is the collection of formal concepts in reference $\mathbf{r}$, $N_r = |\mathcal{C}_r|$, $\mathcal{C}_a$ is a noised collection of bi-sets, $(T_i, G_i)_r \in \mathcal{C}_r$ and $(T_j, G_j)_a \in \mathcal{C}_a$. When $\sigma(\mathcal{C}_r, \mathcal{C}_a) = 1$, all the bi-sets $\in \mathcal{C}_r$ have identical instances in $\mathcal{C}_a$.

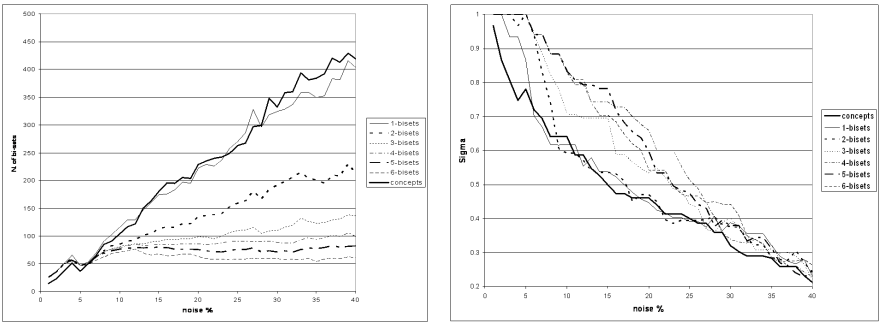

Fig. 1. Size of different collections of bi-sets (left) and related values of σ (right) depending on noise level

In the experiment, **r** has 30 objects and 15 properties and it contains 3 formal concepts of the same size which are pair-wise disjoints. In other terms, the formal concepts are $(\{o_1, \ldots, o_{10}\}, \{p_1, \ldots, p_5\})$, $(\{o_{11}, \ldots, o_{20}\}, \{p_6, \ldots, p_{10}\})$, and $(\{o_{21}, \ldots, o_{30}\}, \{p_{11}, \ldots, p_{15}\})$. We generated 40 different data sets by adding to **r** increasing quantities of noise (from 1% to 40% of the matrix). Then, for each data set, we have extracted a collection of formal concepts and different collections of δ-bi-sets with increasing values of δ (from 1 to 6). Finally, we looked for the occurrence of the 3 formal concepts in each of these extracted collections by using our σ measure. Results are in Fig. 1.

The σ measure decreases when the noise level increases. Interestingly, its values for δ-bi-set collections are always greater or similar to the values for the collection of formal concepts. The collections of δ-bi-sets contain always less patterns than the collection of formal concepts (for a noise level greater than 7%). For $\delta = 2$, the size is halved. For greater values of δ, noise does not influence the size of the collections of δ-bi-sets. This experiment confirms that δ-bi-sets are more robust to noise than formal concepts. Furthermore, it enables to reduce significantly the size of the extracted collections and this is important to support the interpretation process.

3.3 Using Association Rules

Association rules can be derived from extracted bi-sets and used for bi-cluster characterization. For characterization but also classification, heuristics have been studied which select relevant association rules based on their frequency and confidence values [16,17,18]. In our case, we propose to use exception ratios on the extracted bi-sets to provide characterization rules. They have the form $X \Rightarrow v$ where X is a set of properties (resp. objects if the transposed matrix is used) and v is a variable denoting a cluster of objects (resp. properties). When considering formal concepts, deriving characterization rules from them is straightforward.

Property 1. Given a bi-cluster (C^o, C^p), if (T, G) is a formal concept, then $G \Rightarrow C^o$ (resp. $T \Rightarrow C^p$) is a rule with frequency equal to $|T| \cdot (1 - \epsilon_o)$ (resp. $|G| \cdot (1 - \epsilon_p)$) and confidence equal to $1 - \epsilon_o$ (resp. $1 - \epsilon_p$).

When we use δ-bi-sets instead of formal concepts, Property 1 does not hold because $|\psi(G, \mathbf{r})| < |T|$. However, if we are interested in characterizing a cluster of objects, we can use the following property:

Property 2. Given a cluster C^o, if (T, G) is a δ-bi-set, and $X \subseteq G$ is a δ-free-set then $X \Rightarrow C^o$ is a rule with frequency equal to $|T| \cdot (1 - \epsilon_o)$ and confidence equal to $1 - \epsilon_o$.

Such rules are interesting in practice because X is often a rather small set such that its interpretation is easier. However, this approach can not be applied to data sets with large numbers of properties (e.g., for gene expression data sets where thousands of properties are common). In such cases, we propose to use the ϵ_o and ϵ_p measures.

3.4 Examples of Characterizing Queries

So far, we have a methodology for characterizing (bi-)clusters by using different kinds of bi-sets or association rules which can be derived from them. Proposed accuracy measures can be used for a direct selection of characterizing patterns by means of queries:

- Select all the bi-sets which characterize bi-cluster (C^o, C^p) with a maximum exception ratio of ε for both objects and properties;
- Select all the rules with minimal body characterizing bi-cluster (C^o, C^p) with a minimal frequency f, a minimal confidence c, and a maximal exception ratio ε for the set of properties;
- Select all the rules with minimal body characterizing bi-cluster (C^o, C^p) with a minimal frequency f, a minimal confidence c, and a minimal exception ratio ε for the set of properties.

The last example is interesting since it returns bi-sets (or rules) that are exceptions, i.e., they concern objects belonging to bi-cluster (C^o, C^p) that are characterized by some properties from other bi-clusters.

4 Experimental Validation

First, we applied our characterization method to the well-known benchmark voting-records [19]. It contains 435 objects and 48 Boolean attributes (removing class variables). We used Cocluster [6] to get 2 bi-clusters:

bi-cluster	$\lvert\boldsymbol{\tau}\rvert$	rep.	dem.	$\lvert\boldsymbol{\gamma}\rvert$
bi-cluster1	193	153	40	16
bi-cluster2	242	15	227	32
total	435	168	267	48

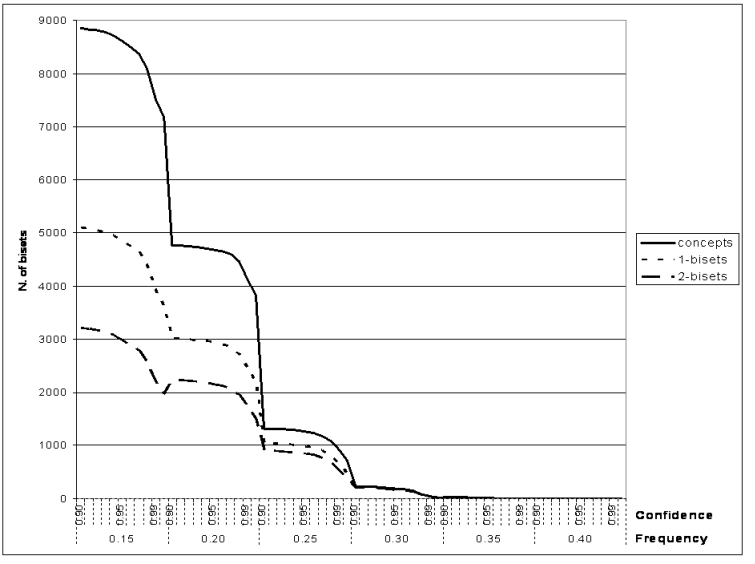

Fig. 2. Characterizing patterns for bi-cluster1 in voting-records w.r.t. different values of minimal frequency and confidence

To characterize each bi-cluster, we used D-Miner [11] to extract all formal concepts, and our slight extension of AcMiner to extract two collections δ-bi-sets (δ=1,2). We obtained 227 031 formal concepts, 130 313 1-bi-sets and 66 908 2-bi-sets. The collections have been post-processed by looking for rules with increasing values of the relative minimal frequency (15% up to 40%) and confidence (90% up to 100%). Results for the first bi-cluster are in Fig. 2. Results for the second one look similar. The number of characterizing rules decreases when we increase the frequency and confidence thresholds. When we use δ-bi-sets, we have to process significantly smaller collections. Two examples of characterizing rules which are consistent with the domain knowledge associated to voting-records are now given. The first one (resp. the second) has a 42% relative frequency (resp. 31%) and both have a 100% confidence, i.e., we have $\epsilon_o = 0$.

```
el-salvador-aid = yes ∧ anti-satellite-test-ban = yes
                ∧ aid-to-nicaraguan-contras = yes ⇒ bi-cluster2
handicapped-infants = no ∧ physician-fee-freeze = yes
                ∧ el-salvador-aid = yes ⇒ bi-cluster1
```

Then, we applied the method to the real world medical data set meningitis already used in [18]. It has been gathered from children hospitalized for acute meningitis. The pre-processed Boolean data set is composed of 329 examples described by 60 Boolean attributes encoding clinical signs (hemodynamic troubles, consciousness troubles, ...), cytochemical analysis of the cerebrospinal fluid (C.S.F proteins, C.S.F glucose, ...), and blood analysis (sedimentation rate,

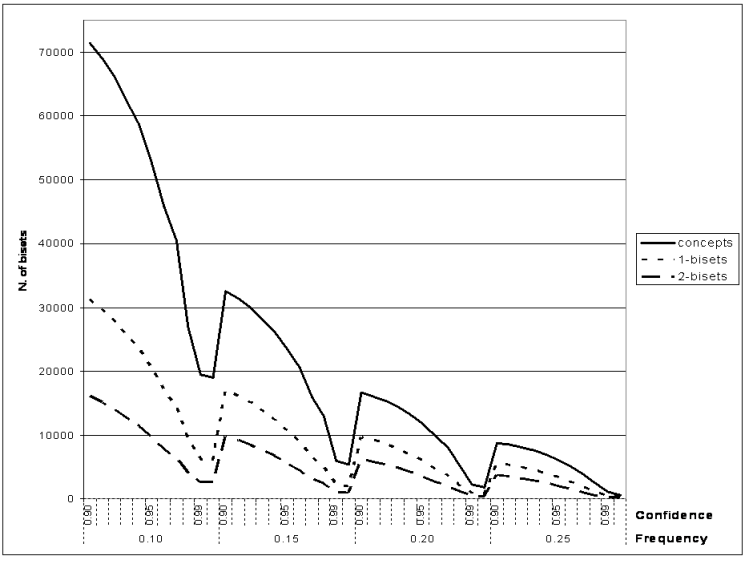

Fig. 3. Characterizing patterns for the bi-cluster2 in meningitis w.r.t. different values of minimal frequency and confidence

white blood cell count, ...). In meningitis, the majority of the cases are known to be viral infections whereas about one quarter are are known to be caused by bacteria. Furthermore, medical knowledge is available which can be used to assess characterization relevancy. Using COCLUSTER, we got two bi-clusters:

bi-cluster	$\|\boldsymbol{\tau}\|$	bact.	vir.	$\|\boldsymbol{\gamma}\|$
bi-cluster1	100	81	19	21
bi-cluster2	229	3	226	39
total	329	84	245	60

The first bi-cluster contains a majority of bacterial cases while the second one contains almost only viral cases. We selected characterization rules based on a collection of formal concepts and 2 collections of δ-bi-sets (δ=1,2). We obtained the results in Fig. 3. Here again, using δ-bi-sets leads to smaller collections of candidate characterization patterns. The number of characterization rules for the first bi-cluster is always very low and it does not significantly change when using δ-bi-sets instead of formal concepts. If we select the rules with a minimal body, a 10% frequency threshold, a 98% confidence threshold, and for which the property exception ratio ϵ_p is zero, we obtain only 9 rules which are consistent with the medical knowledge (see [18] for details). Examples of rules are:

`presence of bacteria in C.S.F. analysis = yes` $\Rightarrow$ `bi-cluster1`
`polynuclear percent` $> 80 \wedge$ `C.S.F. proteins` $> 0.8 \Rightarrow$ `bi-cluster1`
`C.S.F. proteins` $> 0.8 \wedge$ `C.S.F. glucose` $< 1.5 \Rightarrow$ `bi-cluster1`

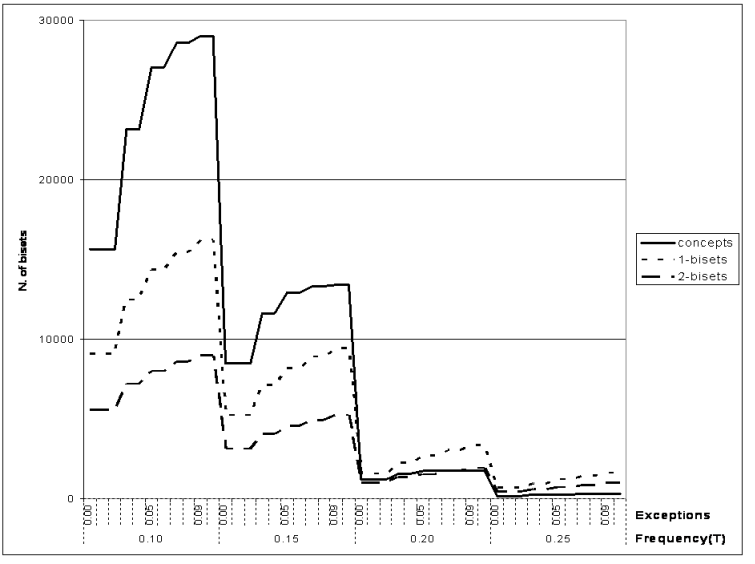

Fig. 4. Characterizing bi-sets for bi-cluster1 in plasmodium w.r.t. different values of minimal size and maximal exception ratio

Finally, our last experiment concerns the analysis of plasmodium, a public gene expression data set concerning Plamodium falciparum (i.e., a causative agent of human malaria) described in [20]. It records the expression profile of 3 719 genes in 46 biological samples. Each sample corresponds to a time point of the developmental cycle. It is divided into 3 phases: the ring, the trophozoite and the schizont stages. The numerical expression data have been preprocessed by using one of the property encoding methods described in [21]. We used Co-cluster to get the following bi-clusters.

bi-cluster	$\|\boldsymbol{\tau}\|$	ring	troph	schiz.	$\|\boldsymbol{\gamma}\|$
bi-cluster1	20	15	5	0	558
bi-cluster2	16	0	5	11	1699
bi-cluster3	10	6	0	4	1462
total	46	21	10	15	3719

We extracted collections of bi-sets to characterize clusters of samples by means of sets of genes. In this case however, the number of properties (columns) was too large to be processed and we extracted the collections of δ-bi-sets on the transposed matrix. Obviously, the frequency and confidence measure do not make sense any more because they are computed on sets of samples and we are looking for sets of genes. Therefore, we have used the size of the bi-sets $|T|$ and $|G|$, and their exception ratios ϵ_o and ϵ_p. Results for a minimal size from 10% up to 25% of $\mathcal{O}$ and for maximal values of ϵ_o from 0% up to 10% are in Fig. 4.

Considering bi-cluster1, we analyzed the characterizing 2-bi-sets whose the minimal size for their sets of objects was 25% of $\mathcal{O}$ and for a maximal exception ratio $\varepsilon_o = 0$. Among the 442 bi-sets characterizing bi-cluster1, only 4 of them concern genes that belong to the same bi-cluster. In each of them, we found at least one gene belonging to the cytoplasmic translation machinery group which is known to be active in the ring stage (see [20] for details), i.e., the majority developmental stage within bi-cluster1.

5 Conclusion

We presented a new (bi-)cluster characterization method based on extracted local patterns, more precisely formal concepts and δ-bi-sets. One motivation is that it is now possible to use quite efficient constraint-based mining techniques for various local patterns and it makes sense to consider their multiple uses. While a bi-partition provides a global and generally expected characterization, selected collections of characterizing bi-sets point out local association which might lead to unexpected but relevant information. We strongly believe in the complementarity between global pattern and local pattern mining techniques when considering the whole knowledge discovery process. Our perspective is now to consider the somehow convergent techniques developed for (conceptual) clustering, subgroup discovery [22], summarization by association rules in order to support real-life knowledge discovery processes in functional genomics.

Acknowledgements. The authors wish to thank P. Francois and B. Crémilleux who provided the data set meningitis. They also thank C. Rigotti, J. Besson and C. Robardet for exciting discussions. This research is partially funded by ACI MD 46 (CNRS STIC 2004-2007) Bingo (Bases de Données Inductives pour la Génomique).

References

1. Jain, A., Dubes, R.: Algorithms for clustering data. Prentice Hall, Englewood cliffs, New Jersey (1988)
2. Eisen, M., Spellman, P., Brown, P., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. PNAS **95** (1998) 14863–14868
3. Fisher, D.H.: Knowledge acquisition via incremental conceptual clustering. Machine Learning **2** (1987) 139–172
4. Cheng, Y., Church, G.M.: Biclustering of expression data. In: Proceedings ISMB 2000, San Diego, USA, AAAI Press (2000) 93–103
5. Robardet, C., Feschet, F.: Efficient local search in conceptual clustering. In: Proceedings DS'01. Volume 2226 of LNCS., Springer-Verlag (2001) 323–335
6. Dhillon, I.S., Mallela, S., Modha, D.S.: Information-theoretic co-clustering. In: Proceedings ACM SIGKDD 2003, Washington, USA, ACM Press (2003) 89–98
7. Madeira, S.C., Oliveira, A.L.: Biclustering algorithms for biological data analysis: A survey. IEEE/ACM Trans. Comput. Biol. Bioinf. **1** (2004) 24–45
8. Wille, R.: Restructuring lattice theory: an approach based on hierarchies of concepts. In Rival, I., ed.: Ordered sets. Reidel (1982) 445–470

9. Boulicaut, J.F., Bykowski, A., Rigotti, C.: Free-sets: a condensed representation of boolean data for the approximation of frequency queries. Data Mining and Knowledge Discovery journal **7** (2003) 5–22
10. Stumme, G., Taouil, R., Bastide, Y., Pasqier, N., Lakhal, L.: Computing iceberg concept lattices with TITANIC. Data & Knowledge Engineering **42** (2002) 189–222
11. Besson, J., Robardet, C., Boulicaut, J.F.: Constraint-based mining of formal concepts in transactional data. In: Proceedings PaKDD'04. Volume 3056 of LNAI., Sydney (Australia), Springer-Verlag (2004) 615–624
12. Besson, J., Robardet, C., Boulicaut, J.F.: Mining formal concepts with a bounded number of exceptions from transactional data. In: Proceedings KDID'04. Volume 3377 of LNCS., Springer-Verlag (2004) 33–45
13. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of ACM SIGMOD'93, Washington, D.C., USA, ACM Press (1993) 207–216
14. Boulicaut, J.F., Bykowski, A., Rigotti, C.: Approximation of frequency queries by mean of free-sets. In: Proceedings PKDD'00. Volume 1910 of LNAI., Lyon, F, Springer-Verlag (2000) 75–85
15. Crémilleux, B., Boulicaut, J.F.: Simplest rules characterizing classes generated by delta-free sets. In: Proceedings ES 2002, Cambridge, UK (2002) 33–46
16. Liu, B., Hsu, W., Ma, Y.: Integrating classification and association rule mining. In: Proceedings KDD'98, New York, NY (1998) 80–86
17. Li, W., Han, J., Pei, J.: CMAR: Accurate and efficient classification based on multiple class-association rules. In: Proceedings ICDM'01, San Jose, CA (2001) 369–376
18. Robardet, C., Crémilleux, B., Boulicaut, J.F.: Characterization of unsupervized clusters by means of the simplest association rules: an application for child's meningitis. In: Proceedings IDAMAP'02 co-located with ECAI'02, Lyon, F (2002) 61–66
19. Blake, C., Merz, C.: UCI repository of machine learning databases (1998)
20. Bozdech, Z., Llinás, M., Pulliam, B.L., Wong, E., Zhu, J., DeRisi, J.: The transcriptome of the intraerythrocytic developmental cycle of plasmodium falciparum. PLoS Biology **1** (2003) 1–16
21. Pensa, R.G., Leschi, C., Besson, J., Boulicaut, J.F.: Assessment of discretization techniques for relevant pattern discovery from gene expression data. In: Proceedings ACM BIOKDD'04, Seattle, USA (2004) 24–30
22. Gamberger, D., Lavrac, N.: Expert-guided subgroup discovery: Methodology and application. JAIR **17** (2002) 501–527