Characterization of unsupervised clusters with the simplest association rules: application for child's meningitis

Céline Robardet[1] and **Bruno Crémillleux**[2] and **Jean-François Boulicaut**[1]

Abstract. We combine different recent data mining techniques to improve the symbolic description of unsupervised clusters. First, we use a clustering method that computes bi-partitions (a partition of examples and a related partition of attribute-value pairs). Then, we use an efficient association rule mining technique to describe the membership of examples within each cluster. We propose a technique for removing rules that are not relevant enough for the cluster characterization. An experimental validation on a real world medical data set is provided.
Keywords. Conceptual clustering, association rule, characterization of clusters.

1 Introduction

Unsupervised clustering algorithms compute a partition of a set of examples into clusters such that examples within a cluster are similar. Recently, an important research effort has been devoted to the integration of cluster characterization into such methods. In conceptual clustering, examples are given by attribute-value pairs (e.g., the definition of medical symptoms). These conceptual clustering methods (see e.g., [Ste87, Fis96, TB01]) associate to each cluster a set of vectors containing the probability of appearance of each attribute-value pair. This interpretation can be improved by assigning to each cluster of examples a cluster of attribute-value pairs such that they are mostly in relation with the examples of the associated cluster. Furthermore, computing the partition of examples can be linked to the computation of the partition for attribute-value pairs. This is the case of the Double Clust method that is used here [RF01a, RF01b]. Unfortunately, this characterization might be not sufficient to understand the likeness of the examples within a given cluster. To support the interactive knowledge discovery process, we think that it is interesting to look for rules that could exhibit strong relations between examples and their clusters.

In this paper, we propose a method to improve the understandability of clusters by means of association rules. First we introduce the clustering algorithm Double Clust and our strategy for association rule mining. Then, we propose an algorithm that reduces the discovered collection of rules to keep only the most general ones. Finally we provide an experimental validation of this method on a real data set concerning child's meningitis.

[1] INSA de Lyon, LISI, Bâtiment Blaise Pascal, F-69621 Villeurbanne cedex. Email. celine.robardet@lisi.insa-lyon.fr
[2] GREYC - CNRS UMR 6072 - Université de Caen, Dept informatique, Campus Cote de Nacre, F-14032 Caen Cedex

2 The clustering method: Double Clust

This method is based on the search of two partitions forming a *bi-partition*: one, called X, divides the set of examples, and the other one, denoted by Y, divides the set of attribute-value pairs. To obtain a useful *bi-partition*, this search is guided by an objective function which consists in building a bijection between the clusters of X and those of Y such that if X_i (a cluster of X) and Y_j (a cluster of Y) are in relation with respect to this bijection, the number of relations between an example of X_i and an attribute-value of Y_j must be maximum. On the other hand, the number of relations between an example of X_i and an attribute-value of Y_l ($l \neq j$) must be minimum.

We illustrate this mechanism with the following example. The table on the left is a data set made of 9 examples (the rows) described by 6 boolean items (the columns, each item is an attribute whose value can be true or false). There is a 1 between an example and an item if the example is in relation with the item. Double Clust aims at defining a partition X of examples and a partition Y of items like the one in the table on the right. In this table, rows and column have been permutated as sketched before.

Items

1	0	0	0	0	0
0	0	0	1	0	1
0	0	1	1	0	0
0	1	0	0	1	1
0	1	0	0	0	1
0	0	1	1	0	0
1	0	0	0	0	0
1	0	0	0	0	0
0	0	0	0	0	0

	Y1		Y2			Y3
X1	1	1	0	0	0	0
	1	0	0	0	1	0
	1	1	0	0	0	0
X2	0	0	1	0	1	0
	0	0	1	1	1	0
X3	0	0	0	0	0	1
	0	0	0	0	0	1
	0	0	0	0	0	1
	0	0	0	0	0	0

Figure 1. An example of dataset and its structuring by a *bi-partition*

In this bi-partition, X_1 contains examples number 3, 2 and 6 (row number) while partition Y_1 contains items A_4 and A_3. Also, X_2={5,4}, X_3={7,8,9,1}, $Y_2=\{A_2, A_5, A_6\}$, and $Y_3=\{A_1\}$.

Two clusters X_i and Y_j, which are in relation with respect to the bijection, can be considered as a "concept" in concept lattices [Wil89]. However, there are two main differences. The first one is that for us each example and each item are only considered within a single concept. The second one is that we accept "impure" concepts and thus cope with noisy data.

The objective function must follow some properties to be adapted to the clustering structure, such as the independence upon cluster permutations or the ability to treat bi-partitions having partitions with different numbers of clusters, etc [RF01a]. These properties are checked by measures which evaluate the link between the two partitions X and Y upon a same set on the basis of a contingency table. In Double Clust, the objective function is the asymmetrical τ measure designed by Goodman and Kruskal [GK54]. It is evaluated on a co-occurrence table (p_{ij}). It has been shown in [RF01a] that this measure discriminate well, even in a noisy context, the set of *bi-partitions* regarding the intensity of the functional link existing between the both partitions. p_{ij} is the frequency of relations between an example of X_i and an attribute-value pair of Y_j and $p_{i.} = \sum_j p_{ij}$. When we try to improve the partition Y of attribute-value pairs with respect of the partition X of examples we use the following measure:

$$\tau_{Y/X} = \frac{\sum_i \sum_j \frac{p_{ij}^2}{p_{i.}} - \sum_j p_{.j}^2}{1 - \sum_j p_{.j}^2}$$

When the partition Y is fixed and we want to improve the X one, we use the same measure on the transposed co-occurrence table (denoted by $\tau_{X/Y}$). The principle of Double Clust algorithm consists in alternatively changing one partition while the other one is unchanged (see Algorithm 1).

Algorithm 1 Double Clust

```
Let (X^0, Y^0) be a random bi-partition
i ← 1
Repeat
    Let Y_i being fixed, we search
        X^{i+1} = argmax_{X ∈ V_{X_i}} τ_{X/Y} (X, Y^i)
    Let X_{i+1} being fixed, we serach
        Y^{i+1} = argmax_{Y ∈ V_{Y_i}} τ_{Y/X} (X^{i+1}, Y)
    i ← i + 1
Until τ_{X/Y}(X, Y) = τ_{Y/X}(X, Y)
```

$\mathcal{V}_{X_i}$ denotes a set of neighboring partitions of X_i, and $\mathcal{V}_{Y_i}$ denotes a set of neighboring partitions of Y_i.

This method has given good experimental results on synthesized as well as on UCI repository data sets [Rob02].

We use Double Clust to define a function $\mathcal{C}$ which assigns to each example e a cluster, denoted by C_i, $i : 1 \ldots k$, $k \geq 2$ the number of clusters. Note that the number of clusters if not fixed beforehand for this method.

3 Understandable set of rules characterizing clusters

3.1 Simplest rules characterizing clusters

Rule induction is an important research area in data mining. One popular data mining technique concerns the discovery of association rules [AIS93]. Association rules can tell something like "It is frequent that when properties A_1 and A_2 are true within an example, then property A_3 tends to be true". We give a simple formalization of this mining task.

Definition 1 (itemset, example) *Assume* $\mathbf{R} = \{A_1, \ldots, A_n\}$*, is a schema of attribute-value pairs denoted by boolean descriptors. One attribute-value pair from* $\mathbf{R}$ *is called an* item *and a subset of* $\mathbf{R}$ *is called an* itemset*.* $\mathbf{r}$*, an instance of* $\mathbf{R}$*, is a multi-set of* examples*.*

Definition 2 (association rule) *Given* $\mathbf{r}$*, an instance of* $\mathbf{R}$*, an* association rule *on* $\mathbf{r}$ *is an expression* $D \Rightarrow B$*, where* $D \subseteq \mathbf{R}$ *and* $B \in \mathbf{R} \setminus D$*.*

Semantics of such rules are captured by the classical measures of *frequency* and *confidence*. The frequency of a rule $RHS \Rightarrow LHS$ is the percentage of examples in $\mathbf{r}$ that support $RHS \cup \{LHS\}$. An example supports an itemset if each item of the itemset is true for the example. The confidence of a rule $RHS \Rightarrow LHS$ is the percentage of examples in $\mathbf{r}$ that support RHS among those supporting LHS. It measures a conditional probability to observe the item RHS true in an example that supports LHS.

The standard association rule mining task concerns the discovery of *every* rule whose frequency and confidence are greater than user-specified thresholds. In other terms, one wants rules that are frequent and valid "enough". The main algorithmic issue concerns the computation of every frequent set (if γ denotes the frequency threshold, an itemset is said *frequent* or γ-frequent, if its frequency is greater or equal to γ). The complexity of frequent itemset mining is exponential with the number of items but a huge research effort has addressed this problem the last 5 years (see, e.g., [AMS+96, Bay98, PBTL99, BBR00]).

Finding rules that characterize clusters can be viewed as a special form of association rule mining where conclusions of rules are pre-specified [LHM98]. Let us call C_i the item associated to the i-th cluster number. To characterize C_i, a naive approach is to extract all rules that conclude on C_i. Nevertheless, in practice, a huge number of rules will be provided such that the user is not able to have a synthetic feedback from them. Furthermore, rule sets contain many redundant rules and some rules are over-fitted. Indeed, the first problem has been well-identified and several propositions have been made for deriving non redundant association rules from closed itemsets (see, e.g., [Zak00]). Recent works revisit these questions and bring improvements. In [JL01], memory consumption and time complexity have been decreased by features selection and, in a post-processing stage, rules covering most examples are selected. CMAR [LHP01] uses statistical techniques to avoid bias and improve efficiency by relevant data structures.

To cope with these drawbacks, we consider the efficient extraction of the set of the simplest rules characterizing a feature (here a cluster). This process is based on the properties of δ-free sets and δ-strong rules introduced in [BBR00]. We started to study the use of δ-strong rules within a classification context in [CB02].

Definition 3 (δ-strong rules) *Given* $\mathbf{r}$*, an instance of* $\mathbf{R}$*, a frequency threshold* γ*, an integer* δ*,* $D \subseteq \mathbf{R}$*, and* $B \in \mathbf{R} \setminus D$*, a* δ-strong rule *on* $\mathbf{r}$ *is an association rule* $D \Rightarrow B$ *with at most* δ *exceptions. In other words, its confidence is at least equal to* $1 - (\delta / (\gamma \times |\mathbf{r}|))$*.*

From a technical perspective, δ-strong rules can be built from δ-free sets that will constitute their left-hand sides. It is out of the scope of this paper to provide details about the concept of δ-free set [BBR00]. It is related to the concepts of closed itemset [PBTL99] and almost-closure [BB00].

A key point for the characterization of clusters is that the δ-strong rule formalism offers a property of *minimal body*.

Definition 4 (rule with a minimal body) *Given a frequency threshold* γ *and an integer* δ*, a rule* $D \Rightarrow B$ *has a* minimal body *if there is no frequent rule* $E \Rightarrow B$ *with* $E \subset D$ *and a confidence greater or equal to* $1 - (\delta / (\gamma \times |\mathbf{r}|))$*.*

This definition means that we consider only shortest itemsets to end up on B, the uncertainty being controlled by δ. Clearly, more frequent and valid (with high enough confidence) rules concluding on B can exist. Nevertheless, we prove in [CB02] that, under a sensible assumption about δ and γ (the intuition is that δ must not be too large with respect to γ), any specified rule $E \Rightarrow B$ characterizes the same cluster as the rule with a minimal body which is included in E. An important result, stemming from properties of δ-free sets [BBR00], is that, if $D \Rightarrow B$ is a rule with a minimal body, then D is a δ-free set. [CB02] describes how to infer all rules with minimal body from the whole collection of δ-free sets. Also, the experiments in [CB02] show that the number of characterizing rules drastically decreases. We call $\mathcal{R}$ the set of rules with a minimal body characterizing clusters.

We argue that this property of minimal body is a fundamental issue for class or cluster characterization. Not only it prevents from overfitting but also it makes the characterization of an example easier to explain. It provides a feedback on the application domain expertise that can be reused for further analysis.

3.2 Selection of rules

3.2.1 Indicators of quality

To improve the understandability of the clustering output, it is useful to keep from $\mathcal{R}$ only the "best" rules for characterization. The following indicators can be used as a basis for measuring the characterization quality. These indicators take into account the quality of prediction of a cluster by the rule set:
$Cover^+$ (resp. $Cover^-$) gives for each example (denoted e) the weighted frequency of correctly (resp. incorrectly) triggered rules of $\mathcal{R}$. Let us denote by $e \models D \Rightarrow B$ the fact that all the items from $D \cup \{B\}$ are true in e and $e \not\models D \Rightarrow B$ the fact that all the items from D are true in e and B is false in e.

$$Cover^+(\mathcal{R}, e) = \sum_{D\Rightarrow c\in\mathcal{R}} \frac{\delta_{e\models D\Rightarrow c}}{\sharp\,(D\Rightarrow c\in\mathcal{R} \mid e\models D\Rightarrow c)} \times fq_{\,D\Rightarrow c} \times cf_{\,D\Rightarrow c}$$

$$Cover^-(\mathcal{R}, e) = \sum_{D\Rightarrow c\in\mathcal{R}} \frac{\delta_{e\not\models D\Rightarrow c}}{\sharp\,(D\Rightarrow c\in\mathcal{R} \mid e\not\models D\Rightarrow c)} \times fq_{\,D\Rightarrow c} \times cf_{\,D\Rightarrow c}$$

where $\delta(condition)$ is the Kronecker measure which equals 1 if $condition$ is satisfied, and 0 otherwise, $\sharp(set)$ denotes the cardinal of set, fq_{rule} is the frequency of $rule$, and cf_{rule} is its confidence.

We now consider an objective function that enable to reduce the set of rules from $\mathcal{R}$ to $\mathcal{R}'$. We need a function that evaluates the variation of the covering quality of each set of rules on the whole set of examples. We can consider that an example is correctly predicted by a set of rules if the majority of rules are correctly triggered for this example. Consequently, we can say that an object e is better covered by $\mathcal{R}'$ than by $\mathcal{R}$ if

$$Cover^+(\mathcal{R}', e) - Cover^-(\mathcal{R}', e) \geq Cover^+(\mathcal{R}, e) - Cover^-(\mathcal{R}, e)$$

and

$$Cover^+(\mathcal{R}', e) > Cover^-(\mathcal{R}', e)$$

We denote by α_e and β_e these two conditions and we use the Kronecker measures δ_{α_e} and δ_{β_e}.

Furthermore, we look for a set of rules $\mathcal{R}'$ such that each example of $\mathbf{r}$ is better described by $\mathcal{R}'$ than by $\mathcal{R}$. Consequently, we define our measure by:

$$\Delta Cover\,(\mathcal{R}', \mathcal{R}) = \sum_{e} \delta_{\alpha_e}\delta_{\beta_e}$$

This measure takes its values in $\mathbb{R}^+$. Depending on the rules removed from $\mathcal{R}$, $\Delta Cover\,(\mathcal{R}', \mathcal{R})$ can increase or decrease. If incorrectly triggered rules for e are removed from $\mathcal{R}'$ then α_e can become true (false if it is not the case). In the contrary, if correctly triggered rules for e are removed from $\mathcal{R}'$ then α_e or β_e can become false.

In the following section, we propose a heuristic algorithm that optimizes this function.

3.2.2 The algorithm - An abstract version

We want to select a set of rules $\mathcal{R}'$ which maximize the function $\Delta Cover\,(\mathcal{R}, \mathcal{R}')$, $\mathcal{R}$ being the original set of rules. For that purpose, let us define two indicators which provide the frequency of examples correctly (resp. incorrectly) supported by each rule:

$$Exact^+(D\Rightarrow c, \mathbf{r}) = \frac{\sharp\,(e\in\mathbf{r} \mid e\models D\Rightarrow c)}{\sharp\,(e\in\mathbf{r} \mid \mathcal{C}(e) = c)}$$

$$Exact^-(D\Rightarrow c, \mathbf{r}) = \frac{\sharp\,(e\in\mathbf{r} \mid e\not\models D\Rightarrow c)}{\sharp\,(e\in\mathbf{r} \mid \mathcal{C}(e) \neq c)}$$

The generality of a rule is evaluated by the following measure:

$$\Delta Exact\,(D\Rightarrow c) = Exact^+(D\Rightarrow c, \mathbf{r}) - Exact^-(D\Rightarrow c, \mathbf{r})$$

We use this function to select the most adequate rules for describing the clustering class. The following heuristic algorithm (Algorithm 2) compute these rules. We note SR and SR_{min} two sets of selected rules.

Algorithm 2 Reducing the collection of characterizing rules

$\mu \leftarrow 0$
$SR_{min} \leftarrow R$
Repeat
  $SR \leftarrow SR_{min}$
  $SR_{min} \leftarrow SR \setminus \{r \in SR \mid \Delta Exact\,(r) \leq \mu\}$
  $\mu \leftarrow \mu + \epsilon$
Until $(\Delta Cover\,(SR_{min}, R) \leq \Delta Cover\,(SR, R))$
Return SR

In this algorithm, less adequate rules w.r.t. a set of examples $\mathbf{r}$, are removed such that a maximum of example clustering class values can be correctly guess.

4 Application

4.1 The data

Let us now describe the results we got by running Double Clust on a real world data set and then extracting the rules for cluster characterization.

The data set is a medical dataset coming from the University of Medicine at Grenoble (France) that concerns child's meningitis. It contains the data gathered from children (whose age was between 10 days and 15 years inclusively) hospitalized for acute meningitis in the pediatric service over a period of 4 years.

Faced with a child with a case of acute meningitis, the clinician must quickly decide which medical course should be taken. Briefly, the majority of these cases are viral infections for which a simple medical supervision is sufficient, whereas about one quarter of the cases are caused by bacteria and need treatment with suitable antibiotics. In typical cases, the diagnosis can be considered as obvious and a few simple rules enable a quasi certain decision. However, nearly one third of the cases are presented with non-typical clinical and biological data: the difficulty of the diagnosis lies in the fact that the observed attributes, considered separately, have little diagnostic signification. Two groups (viral versus bacterial) can be expected.

The used data set is composed of 329 examples described by attributes recording clinical signs (hemodynamic troubles, consciousness troubles, fever, purpura, the season when the infection crops up, recent antibiotic treatment…) cytochemical analysis of the cerebrospinal fluid (C.S.F proteins, C.S.F glucose, white cells count and polynuclear percentage) and blood analysis (the sedimentation rate, the white blood cell counts, the polynuclear neutrophil level and the percentage of immature band cells). We have finally obtained 22 (quantitative or qualitative) attributes. Discretization of quantitative attributes has been done according to thresholds given by experts.

There are decision models based on a numerical synthesis of examples coming from multivariate statistical analysis [FCRD92]. Nevertheless, understandability of such models is quite poor. Physicians can easier manage symbolic interpretations. In the following experimentation, we try to identify some homogeneous sets of patients (without taking into account the etiological diagnosis) and then characterize these clusters by a small set of symbolic descriptions, i.e., rules.

4.2 The results

Double Clust has computed a partition of examples in two clusters X_1 et X_2. In this specific application, we had access to the most important feature characterizing patients (the etiological diagnosis) and Table 1 crosses the two clusters with the diagnosis (let us recall that we did not used this attribute during the clustering).

	X_1	X_2	Sum
Bacterial	6	78	84
Viral	233	12	245
Sum	239	90	329

Table 1. Clusters with respect to etiological diagnosis

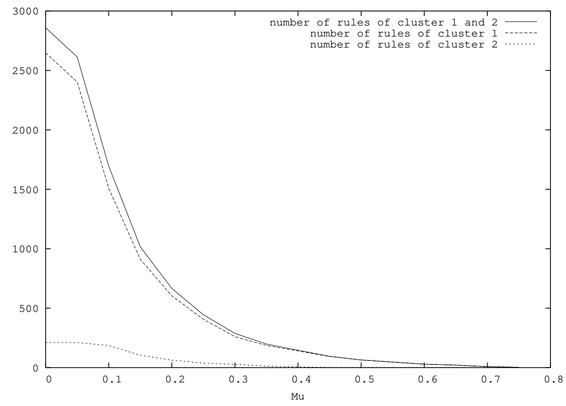

Figure 2. Number of rules at each step of the algorithm

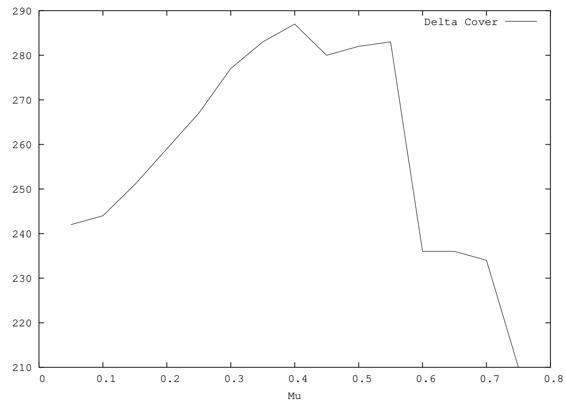

Figure 3. Value of $\Delta Cover$ at each step of the algorithm

We can see that most of children of cluster X_1 suffer from viral meningitis whereas most of the children of cluster X_2 suffer from a bacterial meningitis.

We got 2859 δ-strong rules with minimal body. 2648 conclude on X_1 and 211 on X_2. Obviously, it is difficult for a physician to use such a set of rules. To reduce it and get a better cluster characterization, we applied Algorithm 2 with $\epsilon = 0.05$. This algorithm optimizes the value of the $\Delta Cover$ function by reducing the number of rules in a way which favors most general and correctly triggered rules.

Figure 2 shows the reduction of the number of rules, globally and for each cluster, at each step μ of the algorithm. It stops when there are no more rules. The number of rules of cluster X_1 decreases quickly, whereas the slope for cluster X_2 is slower. This is what we are looking for: removing more rules from the larger (in term of rules) cluster to have the simplest characterization.

Figure 3 shows the value of $\Delta Cover$ at each step of Algorithm 2: it stops when there are no more rules. The shape of the curve is almost convex: it starts with an important increase at the beginning, then we have a quite flat part and it finishes by a fast decrease.

The stop criterion of the algorithm (to stop just before the first decrease) gives a cut point at $\mu = 0.4$. All $\Delta Cover$ values are smaller after this value. Let us note that the number of rules has drastically decreased. Atq this final step, we got 140 rules concluding on X_1 and only 6 on X_2.

Table 2 gives the left-hand side (LHS) of the 6 rules concluding on X_2. All these rules (except the rule number 4 which is difficult to interpret) are consistent with the medical knowledge. We consider that this method brings two improvements. First, the small number of selected rules allows an expert to browse them. Secondly, even when a rule is expected (e.g., the rule with LHS `C.S.F. proteins > 0.8 and C.S.F. glucose < 1.5` is expected), such a method provide objective interestingness measures for the strength of the rule (frequency and confidence values). Notice also that it is easy, at this stage, to compute other interestingness measures based on the frequency (e.g., the lift, the conviction, the J-measure).

Rule	Confidence	Frequency
Presence of bacteria detected in the C.S.F. bacteriological analysis	98.15	16.41
polynuclear percentage $>$ 80 and C.S.F. proteins $>$ 0.8	97.62	12.77
C.S.F. proteins $>$ 0.8 and percentage of immature band cells $<$ 70	95.35	13.07
tonus = unknown and consciousness troubles and no neurological deficiency	97.50	12.16
consciousness troubles and C.S.F. proteins $>$ 0.8	95.00	12.16
C.S.F. proteins $>$ 0.8 and C.S.F. glucose $<$ 1.5	100	10.94

Table 2. Finally selected rules on bacterial meningitis cluster

5 Conclusion

We propose an unsupervised data mining process that combines a clustering method with the generation of a small set of rules that describe the membership of the examples to the clusters. It improves the symbolic characterization of the computed partition by providing for each cluster a small set of general and reliable rules. For that purpose, we propose a function, which compares the characterization quality of two sets of rules, and a heuristic algorithm to select "potentially good" subsets of rules. Our experimentation shows the adequacy of the method on a real data set. The data set is intrinsically composed of two populations: the bacterial and the viral child's meningitis. Those two populations have been clearly identified by the clustering method used. We got two sets of rules, which characterizes the two clusters (bacterial and viral). The comparison of the obtained partition and rules with the available medical knowledge has confirmed the quality of the characterization. We plan to experiment the method on other real-world data sets for further refinement.

Acknowledgements

The authors wish to thank Dr P. Franois (University Hospital of Grenoble, France) for providing data on child's meningitis and many valuable comments.

This research is partially funded by the European Commission IST Programme-Future and Emergent Technologies, under contract cInQ IST-2000-26469.

REFERENCES

[AIS93] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of ACM SIGMOD'93*, pages 207 – 216, Washington, D.C., USA, 1993. ACM Press.

[AMS+96] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and Verkamo A.I. Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*, pages 307 – 328. AAAI Press, 1996.

[Bay98] R. J. Bayardo. Efficiently mining long patterns from databases. In *Proceedings of ACM SIGMOD'98*, pages 85 – 93, Seattle, Washington, USA, 1998. ACM Press.

[BB00] J. F. Boulicaut and A. Bykowski. Frequent closures as a concise representation for binary data mining. In *Proceedings of the Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD'00*, volume 1805 of *Lecture Notes in Artificial Intelligence*, pages 62–73, Kyoto, Japan, 2000. Springer-Verlag.

[BBR00] J. F. Boulicaut, A. Bykowski, and C. Rigotti. Approximation of frequency queries by means of free-sets. In *Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD'00*, volume 1910 of *Lecture Notes in Artificial Intelligence*, pages 75–85, Lyon, F, 2000. Springer-Verlag.

[CB02] B Crémilleux and J. F. Boulicaut. Simplest classification rules generated by δ-free sets. Technical report, INSA Lyon-LISI, 2002. A French preliminar version of this paper has appearead in the proceedings of the French-speaking biennal conference on Artificial Intelligence RFIA'02.

[FCRD92] P. François, B. Crémilleux, C. Robert, and J. Demongeot. MENINGE: a medical consulting system for child's meningitis. study on a series of consecutive cases. *Artificial Intelligence in Medicine*, 4:281–292, 1992.

[Fis96] D. H. Fisher. Iterative optimization and simplification of hierarchical clusterings. *Journal of Artificial Intelligence Research*, 4:147–180, 1996.

[GK54] L. A. Goodman and W. H. Kruskal. Measures of association for cross classification. *Journal of the American Statistical Association*, 49:732–764, 1954.

[JL01] V. Jovanoski and N. Lavrac. Classification rule with apriori-c. In *Proceedings of the Integrating Aspects of Data Mining, Decision Support and Meta Learning workshop, co-located with the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases PKDD 01*, pages 81–92, Freiburg, G, September 2001.

[LHM98] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rules mining. In *Proceedings of Fourth International Conference on Knowledge Discovery & Data Mining, KDD 98*, pages 80–86. AAAI Press, 1998.

[LHP01] W. Li, J. Han, and J. Pei. CMAR: Accurate and efficient classification based on multiple class-association rules. In *proceedings of the IEEE International Conference on Data Mining, ICDM 01*, San Jose, California, 2001.

[PBTL99] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24(1):25–46, Elsevier, 1999.

[RF01a] C. Robardet and F. Feschet. Comparison of three objective functions for conceptual clustering. In L. De Raedt and A. Siebes, editors, *Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases - PKDD 2001*, number 2168 in LNAI, pages 399–410. Springer-Verlag, September 2001.

[RF01b] C. Robardet and F. Feschet. Efficient local search in conceptual clustering. In K. Jantke and A. Shinohara, editors, *Discovery science - DS 2001*, number 2226 in Lecture Notes in Computer Science, pages 323–335. Springer-Verlag, november 2001.

[Rob02] C. Robardet. *Contribution à la classification non supervisée : proposition d'une méthode de bi-partitionnement*. PhD thesis, University Lyon 1 Claude Bernard, F-69622 Villeurbanne cedex, 2002. To appear (in French).

[Ste87] R. Stepp. Concepts in conceptual clustering. In J. P. McDermott, editor, *International Joint Conference on Artificial Intelligence (IJCAI-87)*, pages 211–213. Morgan Kaufmann, august 1987.

[TB01] L. Talavera and J. Béjar. Generality-based conceptual clustering with probabilistic concepts. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):196–206, 2001.

[Wil89] R. Wille. Knowledge acquisition by methods of formal concept analysis. In E. Diday, editor, *Data analysis, learning symbolic and numeric knowledge*, pages 365–380, North-Holland, 1989. Nova Sciences.

[Zak00] M. Zaki. Generating non-redundant association rules. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining SIGKDD'00*, pages 34 – 43, Boston, Massachussets, USA, 2000. AAAI Press.