Optimized Rule Mining Through a Unified Framework for Interestingness Measures

Céline Hébert and Bruno Crémilleux

GREYC, CNRS - UMR 6072, Université de Caen
Campus Côte de Nacre
F-14032 Caen Cédex France
Forename.Surname@info.unicaen.fr

Abstract. The large amount of association rules resulting from a KDD process makes the exploitation of the patterns embedded in the database difficult even impossible. In order to address this problem, various interestingness measures were proposed for selecting the most relevant rules. Nevertheless, the choice of an appropriate measure remains a hard task and the use of several measures may lead to conflicting information. In this paper, we propose a unified framework for a set of interestingness measures $\mathcal{M}$ and prove that most of the usual objective measures behave in a similar way. In the context of classification rules, we show that each measure of $\mathcal{M}$ admits a lower bound on condition that a minimal frequency threshold and a maximal number of exceptions are considered. Furthermore, our framework enables to characterize the whole collection of the rules simultaneously optimizing all the measures of $\mathcal{M}$. We finally provide a method to mine a rule cover of this collection.

1 Introduction

Exploring and analyzing relationships between features is a central task in KDD processes. Agrawal et al. [1] define association rules as the implications $X \rightarrow Y$ where X and Y represent one or several features (or attributes). However, among the overwhelming number of rules resulting from practical applications, it is difficult to determine the most relevant rules [9]. An essential task is to assist the user in selecting interesting rules. In this paper, we focus on classification rules i.e. rules concluding on a class label. These rules are useful for producing emerging patterns [5], characterizing classes [4] and building classifiers [10].

This paper is about the use of interestingness measures for classification rules. Such measures are numerous. The Support and the Confidence measures are probably the most famous ones, but there are more specific ones such as the Lift [3] or the Sebag and Schoenauer's measure [14]. In practice, choosing a suitable measure and determining an appropriate threshold for this measure is a challenge for the end user. Combining results coming from several measures is even much more difficult. Thus an important issue is to compare existing interestingness measures in order to highlight their similarities and differences and understand better their behaviors [13, 2]. Even if most of the usual measures

A Min Tjoa and J. Trujillo (Eds.): DaWaK 2006, LNCS 4081, pp. 238–247, 2006.

are based on the rule antecedent frequency and the rule number of exceptions, there is a lack of generic results and this observation was the starting point of this work.

Contributions. In this paper, we design an original framework grouping together a large set $\mathcal{M}$ of measures (also called δ-dependent measures) having a similar behavior. This framework points out the minimal frequency threshold γ and the maximal number of exceptions of a rule δ as two parameters highly characterizing the rule quality. We provide lower bounds according to γ and δ for any measure in $\mathcal{M}$. This result guarantees a minimal quality for the rules. We show that all the measures $\mathcal{M}$ can be simultaneously optimized, which ensures to produce the best rules according to these measures. We finally provide a method to mine a rule cover of these rules, making our approach efficient in practice.
Organization. The paper is organized as follows. Section 2 discusses related work and gives preliminary definitions. Section 3 describes our framework and the relationship with the parameters γ and δ. Section 4 shows how to optimize the measures $\mathcal{M}$ and obtain a rule cover of all the rules optimizing simultaneously these measures. Section 5 gives experimental results about the quality of the mined rules.

2 Preliminaries

2.1 Covering and Selecting the Most Interesting Rules

Lossless cover. It is well known that the whole set of association rules contains a lot of redundant rules [1]. So several approaches propose to extract a cover of the rules [19] like the *informative rules* [11]. Such a rule has a minimal antecedent and a maximal consequence. They are lossless since we can regenerate the whole set of rules according to both minimal support and confidence thresholds. In Section 4.3, we extend this result by proving that the informative classification rules are a cover of all the rules optimizing simultaneously the measures $\mathcal{M}$.
Deficiencies in selecting the most interesting rules. A lot of works address the selection of relevant rules by means of interestingness measures. It requires to define properties characterizing "good" interestingness measures [9, 12]. Piatetsky-Shapiro [12] proposes a framework with three properties and we set our work with respect to it. Other works compare interestingness measures to determine their differences and similarities, either in an experimental manner [16] or in a theoretical one [15, 7]. There are also attempts to combine several measures to benefit from their joint qualities [6]. However it should be underlined that choosing and using a measure remain a hard task.

This paper deals with the previously mentionned aspects of the rule selection problem. By defining a large set of measures $\mathcal{M}$ behaving in a similar way, choosing one of these measures becomes a secondary issue. We intend to exhibit the minimal properties that a measure must satisfy in order to get the most generic framework. We combine their qualities by showing that they can be simultaneously optimized. To avoid redundancy, we give a method to produce only a cover of the rules optimizing $\mathcal{M}$, i.e. the informative classification rules.

2.2 Definitions

Basic definitions. A database $\mathcal{D}$ is a relation $\mathcal{R}$ between a set $\mathcal{A}$ of *attributes* and a set $\mathcal{O}$ of *objects*: for $a \in \mathcal{A}, o \in \mathcal{O}$, $a\,\mathcal{R}\,o$ if and only if the object o contains the attribute a. A *pattern* is a subset of $\mathcal{A}$. The frequency of a pattern X is the number of objects in $\mathcal{D}$ containing X; it is denoted by $\mathcal{F}(X)$. Let $\mathcal{C} = \{c_1, \ldots, c_n\}$ be a set of class values. Each object in $\mathcal{D}$ is assigned a class label in $\mathcal{C}$. $\mathcal{D}_i$ corresponds to the set of the objects labeled by c_i. Table 1 shows an example of database containing 8 objects and two labels c_1 and c_2.

Table 1. An example of database $\mathcal{D}$

$\mathcal{D}$	Attributes								Classes		
Objects	A	B	C	D	E	F	G	H	c_1	c_2	
o_1	1	0	1	0	1	0	1	0	1	0	
o_2	0	1	1	0	1	0	1	1	1	0	
o_3	1	0	1	0	1	0	0	1	1	0	$\mathcal{D}_1$
o_4	1	0	1	0	1	0	0	1	1	0	
o_5	0	1	1	0	1	1	0	0	0	1	
o_6	1	0	0	1	0	1	0	1	0	1	
o_7	0	1	1	0	1	1	0	1	0	1	$\mathcal{D}_2$
o_8	1	0	1	0	0	1	0	1	0	1	

Classification rules. A rule $r : X \rightarrow c_i$ where X is a pattern and c_i a class label is a *classification rule*. X is the *antecedent* of r and c_i its *consequence*. $\mathcal{F}(Xc_i)$ is the frequency of r and $\mathcal{F}(X)$ the frequency of its antecedent. For instance, $r_1 : F \rightarrow c_2$ and $r_2 : EH \rightarrow c_1$ are classification rules in $\mathcal{D}$ (cf. Table 1). $\mathcal{F}(X) - \mathcal{F}(Xc_i)$ is the *number of exceptions* of r, i.e., the number of objects containing X which are not labeled by c_i. The rule r_2 admits 1 exception (object o_7) and the frequency of its antecedent is equal to 4.
Evaluating objective measures. An interestingness measure is a function which assigns a value to a rule according to its quality. We recall here the well-known Piatetsky-Shapiro's properties [12] which aim at specifying what a "good" measure is. As this paper focuses on classification rules, we formulate them in this context.

Definition 1 (Piatetsky-Shapiro's properties). *Let $r : X \rightarrow c_i$ be a classification rule and M an interestingness measure.*

- *P1: $M(r) = 0$ if X and c_i are statistically independent;*
- *P2: When $\mathcal{F}(X)$ and $|\mathcal{D}_i|$ remain unchanged, $M(r)$ monotonically increases with $\mathcal{F}(Xc_i)$;*
- *P3: When $\mathcal{F}(Xc_i)$ and $\mathcal{F}(X)$ (resp. $|\mathcal{D}_i|$) remain unchanged, $M(r)$ monotonically decreases with $|\mathcal{D}_i|$ (resp. $\mathcal{F}(X)$).*

P2 ensures the increase of M with the rule frequency. Most of the usual measures satisfy P2 (e.g., support, confidence, interest, conviction). However, there are a few exceptions (e.g., J-measure, Goodman-Kruskal, Gini index). In the next section, we will use P2 to define our framework.

3 A Unified Framework for Objective Measures

This section defines our framework gathering various measures in a set $\mathcal{M}$. The key idea is to express a measure according to two parameters: the minimal frequency γ for the rule antecedent and the maximal number of rule exceptions δ. Then, we formalize the influence of δ by associating a δ-dependent function to each measure.

3.1 Dependency on the Parameter δ

Definition 2 indicates that a measure M can be expressed as a two-variable function of the rule frequency and the frequency of its antecedent.

Definition 2 (Associated Function). *Let M be an interestingness measure and $r : X \rightarrow c_i$ a classification rule. $\Psi_M(x, y)$ is the function associated to M i.e. it is equal to $M(r)$ where $x = \mathcal{F}(X)$ and $y = \mathcal{F}(Xc_i)$.*

For instance, when M is the Growth Rate, we obtain: $\Psi_{GR}(x, y) = \frac{y}{x-y} \times \frac{|\mathcal{D} \setminus \mathcal{D}_i|}{|\mathcal{D}_i|}$. Using Ψ_M, the Piatetsky-Shapiro's properties P2 and P3 (cf. Section 2.2) can be formulated as: "Ψ_M monotonically increases with y" and "Ψ_M monotonically decreases with x".

Figure 1 plots $\mathcal{F}(Xc_i)$ according to $\mathcal{F}(X)$ for a classification rule $r : X \rightarrow c_i$. The gray area depicts the condition imposed by γ on the rule antecedent. The hatched area of width δ illustrates the link between the variables $x = \mathcal{F}(X)$ and $y = \mathcal{F}(Xc_i)$ of Ψ_M: bounding the rule number of exceptions by δ ensures that $\mathcal{F}(Xc_i)$ is close to $\mathcal{F}(X)$.

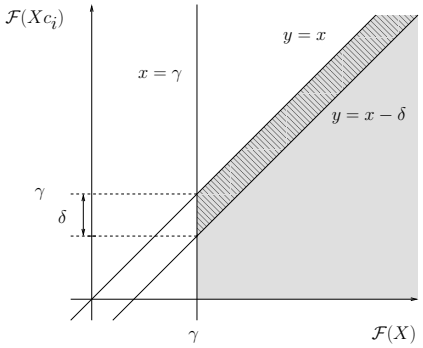

Fig. 1. Dependency between $x = \mathcal{F}(X)$ and $y = \mathcal{F}(Xc_i)$

Definition 3 explicitly expresses this link by defining the now called δ-dependent function $\Psi_{M,\delta}$.

Definition 3 (δ-dependent function). *Let M be an interestingness measure. The δ-*dependent function *$\Psi_{M,\delta}$ is the one-variable function obtained by the change of variable $y = x - \delta$ in Ψ_M i.e., $\Psi_{M,\delta}(x) = \Psi_M(x, x - \delta)$.*

Pursuing the example of the Growth Rate, we get: $\Psi_{GR,\delta}(x) = \frac{x-\delta}{\delta} \times \frac{|\mathcal{D}| - |\mathcal{D}_i|}{|\mathcal{D}_i|}$

3.2 The Set $\mathcal{M}$ of δ-Dependent Measures

We think that the link between $\mathcal{F}(Xc_i)$ and $\mathcal{F}(X)$ highlighted above is an important feature when studying the behavior of an interestingness measure M. That is the reason why we propose a new Property P4 which takes this link into account. P4 imposes that M increases with the variable x of $\Psi_{M,\delta}$.

Definition 4 (P4 : Property of δ-dependent growth). *Let M be an interestingness measure. When δ remains unchanged, $\Psi_{M,\delta}$ increases with x.*

Table 2. Examples of δ-Dependent Measures

Measure	Definition for classification rules	Lower bound
Support	$\frac{\mathcal{F}(Xc_i)}{\lvert\mathcal{D}\rvert}$	$\frac{\gamma-\delta}{\lvert\mathcal{D}\rvert}$
Confidence	$\frac{\mathcal{F}(Xc_i)}{\mathcal{F}(X)}$	$1-\frac{\delta}{\gamma}$
Sensitivity	$\frac{\mathcal{F}(Xc_i)}{\lvert\mathcal{D}_i\rvert}$	$\frac{\gamma-\delta}{\lvert\mathcal{D}_i\rvert}$
Specificity	$1-\frac{\mathcal{F}(X)-\mathcal{F}(Xc_i)}{\lvert\mathcal{D}\rvert-\lvert\mathcal{D}_i\rvert}$	$1-\frac{\delta}{\lvert\mathcal{D}\rvert-\lvert\mathcal{D}_i\rvert}$
Success Rate	$\frac{\lvert\mathcal{D}\rvert-\lvert\mathcal{D}_i\rvert-\mathcal{F}(X)+2\mathcal{F}(Xc_i)}{\lvert\mathcal{D}\rvert}$	$1+\frac{\gamma-2\delta-\lvert\mathcal{D}_i\rvert}{\lvert\mathcal{D}\rvert}$
Lift	$\frac{\lvert\mathcal{D}\rvert\times\mathcal{F}(Xc_i)}{\lvert\mathcal{D}_i\rvert\times\mathcal{F}(X)}$	$(1-\frac{\delta}{\gamma})\times\frac{\lvert\mathcal{D}\rvert}{\lvert\mathcal{D}_i\rvert}$
Piatetsky-Shapiro	$\frac{\mathcal{F}(Xc_i)}{\lvert\mathcal{D}\rvert}-\frac{\lvert\mathcal{D}_i\rvert\times\mathcal{F}(X)}{\lvert\mathcal{D}\rvert^2}$	$[\gamma\times(1-\frac{\lvert\mathcal{D}_i\rvert}{\lvert\mathcal{D}\rvert})-\delta]\times\frac{1}{\lvert\mathcal{D}\rvert}$
Laplace (k=2)	$\frac{\mathcal{F}(Xc_i)+1}{\mathcal{F}(X)+2}$	$\frac{\gamma-\delta+1}{\gamma+2}$
Odds ratio	$\frac{\mathcal{F}(Xc_i)}{\mathcal{F}(X)-\mathcal{F}(Xc_i)}\times\frac{\lvert\mathcal{D}\rvert-\lvert\mathcal{D}_i\rvert-\mathcal{F}(X)+\mathcal{F}(Xc_i)}{\lvert\mathcal{D}_i\rvert-\mathcal{F}(Xc_i)}$	$[\frac{\gamma-\delta}{\lvert\mathcal{D}_i\rvert-\gamma+\delta}]\times[\frac{\lvert\mathcal{D}\rvert-\lvert\mathcal{D}_i\rvert-\delta}{\delta}]$
Growth rate	$\frac{\mathcal{F}(Xc_i)}{\mathcal{F}(X)-\mathcal{F}(Xc_i)}\times\frac{\lvert\mathcal{D}\rvert-\lvert\mathcal{D}_i\rvert}{\lvert\mathcal{D}_i\rvert}$	$\frac{\gamma-\delta}{\delta}\times\frac{\lvert\mathcal{D}\rvert-\lvert\mathcal{D}_i\rvert}{\lvert\mathcal{D}_i\rvert}$
Sebag & Schoenauer	$\frac{\mathcal{F}(Xc_i)}{\mathcal{F}(X)-\mathcal{F}(Xc_i)}$	$\frac{\gamma-\delta}{\delta}$
Jaccard	$\frac{\mathcal{F}(Xc_i)}{\lvert\mathcal{D}_i\rvert+\mathcal{F}(X)-\mathcal{F}(Xc_i)}$	$\frac{\gamma-\delta}{\lvert\mathcal{D}_i\rvert+\delta}$
Conviction	$\frac{\lvert\mathcal{D}\rvert-\lvert\mathcal{D}_i\rvert}{\lvert\mathcal{D}\rvert}\times\frac{\mathcal{F}(X)}{\mathcal{F}(X)-\mathcal{F}(Xc_i)}$	$\frac{\lvert\mathcal{D}\rvert-\lvert\mathcal{D}_i\rvert}{\lvert\mathcal{D}\rvert}\times\frac{\gamma}{\delta}$
ϕ-coefficient	$\frac{\lvert\mathcal{D}\rvert\times\mathcal{F}(Xc_i)-\lvert\mathcal{D}_i\rvert\times\mathcal{F}(X)}{\sqrt{\mathcal{F}(X)\times\lvert\mathcal{D}_i\rvert\times(\lvert\mathcal{D}\rvert-\mathcal{F}(X))\times(\lvert\mathcal{D}\rvert-\lvert\mathcal{D}_i\rvert)}}$	$\frac{\gamma\times(\lvert\mathcal{D}\rvert-\lvert\mathcal{D}_i\rvert)-\delta\times\lvert\mathcal{D}\rvert}{\sqrt{\gamma\times(\lvert\mathcal{D}\rvert-\gamma)\times\lvert\mathcal{D}_i\rvert\times(\lvert\mathcal{D}\rvert-\lvert\mathcal{D}_i\rvert)}}$
Added Value	$\frac{\mathcal{F}(Xc_i)}{\mathcal{F}(X)}-\frac{\lvert\mathcal{D}_i\rvert}{\lvert\mathcal{D}\rvert}$	$\frac{\gamma-\delta}{\gamma}-\frac{\lvert\mathcal{D}_i\rvert}{\lvert\mathcal{D}\rvert}$
Certainty Factor	$\frac{\mathcal{F}(Xc_i)\times\lvert\mathcal{D}\rvert-\mathcal{F}(X)\times\lvert\mathcal{D}_i\rvert}{\mathcal{F}(X)\times(\lvert\mathcal{D}\rvert-\lvert\mathcal{D}_i\rvert)}$	$\frac{\gamma\times(\lvert\mathcal{D}\rvert-\lvert\mathcal{D}_i\rvert)-\delta\times\lvert\mathcal{D}\rvert}{\gamma\times(\lvert\mathcal{D}\rvert-\lvert\mathcal{D}_i\rvert)}$
Information Gain	$\log\left(\frac{\mathcal{F}(Xc_i)}{\mathcal{F}(X)}\times\frac{\lvert\mathcal{D}\rvert}{\lvert\mathcal{D}_i\rvert}\right)$	$\log\left(\frac{\gamma-\delta}{\gamma}\times\frac{\lvert\mathcal{D}\rvert}{\lvert\mathcal{D}_i\rvert}\right)$

We claim that P4 captures an important characteristic of an interestingness measure M: the behavior of M with respect to the joint development of the rule antecedent frequency and the rule number of exceptions. This characteristic is not found in Piatetsky-Shapiro's framework. Table 2 provides a sample of measures fulfilling P4 (see their definitions in [15, 17]).

We define now the set $\mathcal{M}$ of δ-dependent measures:

Definition 5 ($\mathcal{M}$: Set of δ-dependent measures). *The* set of δ-dependent measures $\mathcal{M}$ *is the set of measures satisfying P2 and P4.*

Definition 5 does not require a δ-dependent measure to also satisfy P1 or P3. Remember that our aim is to define the most generic framework. Contrary to P2 and P3, P4 does not impose on $\Psi_{M,\delta}$ to monotonically increase and thus, the Conviction is a δ-dependent measure. A lot of measures belong to $\mathcal{M}$: for instance, all the measures in Table 2 are in $\mathcal{M}$. As $\mathcal{M}$ is defined in intension, it is an infinite set. The next section shows how to bound and optimize these measures.

4 Simultaneous Optimization of the δ-Dependent Measures

4.1 Lower Bounds and Optimization

For any measure of $\mathcal{M}$, Theorem 1 explicits a lower bound depending on γ and δ ($|\mathcal{D}|$ and $|\mathcal{D}_i|$ are constant values).

Theorem 1 (Lower bounds). *Let $r : X \rightarrow c_i$ be a classification rule. Assume that $\mathcal{F}(X) \geq \gamma$ and r admits less than δ exceptions. Thus, for each measure M in $\mathcal{M}$, $\Psi_{M,\delta}(\gamma)$ is a lower bound of $M(r)$.*

Proof. Since r has less than δ exceptions, we immediately have $\mathcal{F}(Xc_i) \geq \mathcal{F}(X) - \delta$. From P2, $\Psi_M(x, y)$ increases with y and thus $\Psi_M(x, y) \geq \Psi_M(x, x - \delta) = \Psi_{M,\delta}(x)$. We know that x is greater than or equal to γ and M satisfies P4. Consequently, $\Psi_{M,\delta}(\gamma)$ is a lower bound for $\Psi_{M,\delta}(x)$. □

Theorem 1 means that the quality of any rule r whose antecedent is a γ-frequent pattern and having less than δ exceptions, is greater than or equal to $\Psi_{M,\delta}(\gamma)$. As this result is true for any measure in $\mathcal{M}$, r satisfies a minimal quality for each measure of $\mathcal{M}$, thus we get a set of rules of good quality according to $\mathcal{M}$. The lower bounds only depends on δ and γ and constant values on $\mathcal{D}$. They can be computed (see the last column in Table 2) to quantify the minimal quality of rules. Intuitively, the more the antecedent frequencies increase and the numbers of exceptions decrease, the higher the global quality of a set of rules is. Due to the properties P2 and P4 of the δ-dependent measures, for all measures M in $\mathcal{M}$, $\Psi_{M,\delta}(\gamma)$ increases with γ and decreases with δ. Property 1 (proved in [8]) shows that the lower bound of any measure M in $\mathcal{M}$ can tend towards its upper bound. The conjunction of Theorem 1 and Property 1 proves that the rules with a γ-frequent antecedent and having less than δ exceptions optimize all the measures in $\mathcal{M}$.

Property 1 *The optimal value of $\Psi_{M,\delta}(\gamma)$ is obtained when $\gamma \rightarrow |\mathcal{D}|$ and $\delta \rightarrow 0$.*

4.2 Completeness

Theorem 2 indicates that any classification rule r optimizing the set of measures $\mathcal{M}$ (i.e., $\forall M \in \mathcal{M},\ M(r) \geq \Psi_{M,\delta}(\gamma)$) is a rule whose antecedent is frequent and having less than δ exceptions.

Theorem 2 (Completeness). *Let $r : X \rightarrow c_i$ a classification rule. Assume that $M(r) \geq \Psi_{M,\delta}(\gamma)$ for all measures M in $\mathcal{M}$. Then $\mathcal{F}(X) \geq \gamma$ and r admits less than δ exceptions.*

Proof. We prove the completeness by reducing it to the absurd. We denote by *Spe* the Specificity and by *Sup* the Support (see Table 2 for their definitions). Assume that r admits δ' exceptions with $\delta' > \delta$. We have $-\frac{\delta'}{|\mathcal{D}|-|\mathcal{D}_i|} < -\frac{\delta}{|\mathcal{D}|-|\mathcal{D}_i|}$

followed by $Spe(r) < \Psi_{Spe,\delta}(\gamma)$, which is in contradiction with our hypothesis. Thus r has less than δ exceptions.

Suppose that $\mathcal{F}(X) = \gamma' < \gamma$. We have $Sup(r) < \frac{\gamma' - \delta}{|\mathcal{D}|} < \Psi_{Sup,\delta}(\gamma)$ and this is in contradiction with our hypothesis as well. Thus, the antecedent frequency is greater than γ. □

The assumption $M(r) \geq \Psi_{M,\delta}(\gamma)$ for $M \in \{Support, Specificity\}$ is sufficient to establish this proof. However, it is necessary to assume $M(r) \geq \Psi_{M,\delta}(\gamma)$ for all measures M in $\mathcal{M}$ in Theorem 2 to demonstrate the completeness of our approach. So, this theorem is the reverse of Theorem 1. Combining Theorems 1 and 2 results in an equality between the set of classification rules with a γ-frequent antecedent and a number of exceptions under δ and the set of classification rules r whose value is greater than or equal to the lower bounds for any measure M in $\mathcal{M}$. Theorem 2 ensures the completeness when mining the rules optimizing the set $\mathcal{M}$ by extracting the classification rules according to the thresholds γ and δ. The next section shows that we can reduce even more the set of rules to mine without any loss.

4.3 Reduction to a Rule Cover

We have introduced in Section 2.1 the rule cover based on informative rules. This cover enables to restore the whole collection of association rules with their exact frequencies and confidence [11]. In this section, we extend this result for the classification rules optimizing simultaneously the measures $\mathcal{M}$. By analogy with informative rules, we call *informative classification rule* a classification rule having a free[1] pattern as antecedent and concluding on a class label. Theorem 3 proves that the informative classification rules having γ-frequent antecedents and less than δ exceptions constitute a cover of the classification rules optimizing all the measures in $\mathcal{M}$.

Theorem 3 (Rule cover). *The set of informative classification rules having γ-frequent antecedents and less than δ exceptions enables to generate the whole set of classification rules r with $M(r) \geq \Psi_{M,\delta}(\gamma)$ for each measure M in $\mathcal{M}$.*

Proof. Assume that $X \rightarrow c_i$ is an informative classification rule and $h(X)$ is the closure of X (see [18] for a definition of the closure). From X and $h(X)$, it is possible to build the set of patterns Y containing X and included in $h(X)$. The rules $Y \rightarrow c_i$ constitute the whole set of rules optimizing $\mathcal{M}$. We demonstrate that:
(1) *A rule $Y \rightarrow c_i$ optimizes all the measures of $\mathcal{M}$.* Due to the properties of the closure, $\mathcal{F}(Y) = \mathcal{F}(X) = \mathcal{F}(h(X))$. Thus, Y is γ-frequent. Moreover, since X and Y appear in the same objects of $\mathcal{D}$, we have $\mathcal{F}(Xc_i) = \mathcal{F}(Yc_i)$. This ensures that the rules $X \rightarrow c_i$ and $Y \rightarrow c_i$ have an identical number of exceptions i.e. less than δ. This proves the first point.

[1] Free (or key) patterns are defined in [11]. They have interesting properties of minimality in lattices and enable to build rules with minimal antecedents.

(2) *All the rules optimizing $\mathcal{M}$ are generated.* Suppose the rule $Z \rightarrow c_i$ optimizes $\mathcal{M}$ but is not generated by our method. Let X' be the largest free pattern containing Z. $X' \rightarrow c_i$ is not an informative rule (otherwise, $Z \rightarrow c_i$ would have been generated) thus we consider two cases: either X' is not frequent but this implies that Z is not frequent (Contradiction) or $X' \rightarrow c_i$ has more than δ exceptions and thus, $Z \rightarrow c_i$ has more than δ exceptions as well (Contradiction). □

For any classification rule r, the cover always contains a rule having the same quality as r for all the measures in $\mathcal{M}$. As there are efficient algorithms to extract the free (or key) patterns [11] which are the antecedents of the informative rules, Theorem 3 is precious to mine in practice the informative classification rules. We have designed the ClaRminer prototype [8] which produces the whole set of informative classification rules.

5 Rule Quality Testing

The aim of the experiments is twofold. We quantify the quality of the rules mined in practice according to the set of measures given in Table 2 and the reduction brought by the informative classification rule cover. Experiments are carried out on the `MUSHROOM` data set from the UCI Machine Learning Repository[2] with a 2.20 GHz Pentium IV processor with Linux operating system by using 3Gb of RAM memory.

Overview of the mined rules. We focus on the quality of the informative classification rules with $\gamma = 812$ and $\delta = 40$. These values correspond to a relative frequency of 10% and a relative number of exceptions under 5%. The mining produces 1598 rules with antecedents containing a maximum of 7 attributes. Table 3 gives for each measure: its lower bound, the average value and the ratio $\frac{avg - \Psi_{M,\delta}(\gamma)}{\Psi_{M,\delta}(\gamma)}$ (called *difference* in Table 3). For instance, the average value for the set of rules is 0.252 for the Sensitivity (the lower bound is 0.184) and 74.65 for the Sebag & Schoenauer's measure (the lower bound is 19.3).For the Sebag & Schoenauer's measure, the average value is 286.81% above the lower bound. Remark that the difference is less important for other measures but the lower bound for these measures is really close to their maximum. For instance, the difference is 4.3% for the Confidence and the lower bound is 0.951 with an optimal value equal to 1.

Table 3. Lower bound, average value and difference (%)

Measure	Support	Confidence	Sensitivity	Specificity	Success Rate	Lift	PS	Laplace	Odds Ratio
Lower bound	0.095	0.951	0.184	0.990	0.572	1.835	0.043	0.950	21.771
Average	0.128	0.992	0.252	0.998	0.620	1.958	0.063	0.991	29.603
Difference	34.74	4.3	37.33	0.81	8.39	6.70	45.63	4.3	35.97

Measure	GR	S & S	Jaccard	Conviction	ϕ-coefficient	AV	Certainty Factor	GI	
Lower bound	17.961	19.30	0.185	9.785	0.289	0.433	0.898	0.607	
Average	71.178	74.654	0.252	36.901	0.72	0.485	0.984	0.671	
Difference	296.29	286.81	36.22	277.12	28.72	12.01	9.58	10.54	

[2] `http://www.ics.uci.edu/~mlearn/MLSummary.html`

Table 4. Ratio number of rules/number of informative rules

γ	812	812	812	1500	1500	1500	2000	2000	2000
δ	25	50	100	50	100	200	75	150	300
Ratio	178.8	143.0	110.5	108.4	64.2	50.8	6.8	6.3	8.9

Comparison between informative rules and the whole set of rules. Table 4 shows that considering only informative rules instead of all the rules optimizing the measures $\mathcal{M}$ significantly reduces the number of rules: according to γ and δ, the cover only contains from 0.6% to 15.9% of the whole rule collection.

6 Conclusions and Future Work

In this paper, we have designed an original framework gathering most of the usual interestingness measures for classification rules. The measures belonging to this framework are shown to behave the same way and choosing "the" appropriate measure appears to be a secondary issue. We have established that all the measures of this framework can be simultaneously optimized, thus enabling the production of the best rules. A cover of these rules can be efficiently mined and experiments indicate that the number of produced rules is significantly reduced.

This work could be extended in many directions. We are working on generalizing our framework to any association rule. One key point of our approach is that the rule consequence is a class label and its frequency is known. The generalization is not obvious because *no information is provided* about the consequence frequency when considering any association rule. Another objective is to automatically determine the couples of parameters (δ, γ) to mine the rules satisfying measure thresholds fixed by the user to combine the various semantics conveyed by the measures. A third extension is the study of measures that do not belong to our framework.

Acknowledgements. This work has been partially funded by the ACI "masse de données" (French Ministry of research), Bingo project (MD 46, 2004-2007).

References

[1] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *SIGMOD'93 Conference*, pages 207–216. ACM Press, 1993.

[2] J. R. J. Bayardo and R. Agrawal. Mining the most interesting rules. In *KDD'99*, pages 145–154, 1999.

[3] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In J. Peckham, editor, *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA*, pages 265–276. ACM Press, 1997.

[4] B. Crémilleux and J.-F. Boulicaut. Simplest rules characterizing classes generated by delta-free sets. In *22nd Int. Conf. on Knowledge Based Systems and Applied Artificial Intelligence (ES'02)*, pages 33–46, Cambridge, UK, December 2002.

[5] G. Dong and J. Li. Efficient mining of emerging patterns: discovering trends and differences. In *proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD'99)*, pages 43–52, San Diego, CA, 1999. ACM Press.

[6] D. Francisci and M. Collard. Multi-criteria evaluation of interesting dependencies according to a data mining approach. In *Congress on Evolutionary Computation*, pages 1568–1574, Canberra, Australia, 12 2003. IEEE Press,.

[7] J. Fürnkranz and P. A. Flach. Roc 'n' rule learning-towards a better understanding of covering algorithms. *Machine Learning*, 58(1):39–77, 2005.

[8] C. Hébert and B. Crémilleux. Obtention de règles optimisant un ensemble de mesures. In *Conférence francophone sur l'apprentissage automatique (CAp'06)*, Trégastel, France, 2006.

[9] R. J. Hilderman and H. J. Hamilton. Measuring the interestingness of discovered knowledge: A principled approach. *Intell. Data Anal.*, 7(4):347–382, 2003.

[10] 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, New York, August 1998. AAAI Press.

[11] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In *proceedings of 7th International Conference on Database Theory (ICDT'99)*, volume 1331 of *Lecture notes in artificial intelligence*, pages 299–312, Jerusalem, Israel, 1999. Springer Verlag.

[12] G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press, 1991.

[13] M. Plasse, N. Niang, G. Saporta, and L. Leblond. Une comparaison de certains indices de pertinence des règles d'association. In G. Ritschard and C. Djeraba, editors, *EGC*, volume RNTI-E-6 of *Revue des Nouvelles Technologies de l'Information*, pages 561–568. Cépaduès-Éditions, 2006.

[14] M. Sebag and M. Schoenauer. Generation of rules with certainty and con dence factors from incomplete and incoherent learning bases. In M. L. J. Boose, B. Gaines, editor, *European Knowledge Acquisistion Workshop, EKAW'88*, pages 28–1–28–20, 1988.

[15] P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *KDD*, pages 32–41. ACM, 2002.

[16] B. Vaillant, P. Lenca, and S. Lallich. A clustering of interestingness measures. In *The 7th International Conference on Discovery Science*, pages 290–297, 10 2004.

[17] B. Vaillant, P. Meyer, E. Prudhomme, S. Lallich, P. Lenca, and S. Bigaret. Mesurer l'intérêt des règles d'association. In *EGC*, pages 421–426, 2005.

[18] R. Wille. *Ordered sets*, chapter Restructuring lattice theory: an approach based on hierachies of concepts, pages 445–470. Reidel, Dordrecht, 1982.

[19] M. J. Zaki. Generating non-redundant association rules. In *KDD'00*, pages 34–43, 2000.