A Parameter-Free Associative Classification Method

Loïc Cerf[1], Dominique Gay[2], Nazha Selmaoui[2], and Jean-François Boulicaut[1]

[1] INSA-Lyon, LIRIS CNRS UMR5205,
F-69621 Villeurbanne, France
{loic.cerf,jean-francois.boulicaut}@liris.cnrs.fr
[2] Université de la Nouvelle-Calédonie, ERIM EA 3791,
98800 Nouméa, Nouvelle-Calédonie
{dominique.gay,nazha.selmaoui}@univ-nc.nc

Abstract. In many application domains, classification tasks have to tackle multiclass imbalanced training sets. We have been looking for a CBA approach (Classification Based on Association rules) in such difficult contexts. Actually, most of the CBA-like methods are one-vs-all approaches (OVA), i.e., selected rules characterize a class with what is relevant for this class and irrelevant for the union of the other classes. Instead, our method considers that a rule has to be relevant for one class and irrelevant for every other class taken separately. Furthermore, a constrained hill climbing strategy spares users tuning parameters and/or spending time in tedious post-processing phases. Our approach is empirically validated on various benchmark data sets.

Key words: Classification, Association Rules, Parameter Tuning, Multiclass

1 Introduction

Association rule mining [1] has been applied not only for descriptive tasks but also for supervised classification based on labeled transactional data [2–8]. An association rule is an implication of the form $X \Rightarrow Y$ where X and Y are different sets of Boolean attributes (also called items). When Y denotes a single class value, it is possible to look at the predictive power of such a rule: when the conjunction X is observed, it is sensible to predict that the class value Y is true. Such a shift between descriptive and predictive tasks needs for careful selection strategies [9]. [2] identified it as an associative classification approach (also denoted CBA-like methods thanks to the name chosen in [2]). The pioneering proposal in [2] is based on the classical objective interestingness measures for association rules – *frequency* and *confidence* – for selecting candidate classification rules. Since then, the selection procedure has been improved leading to various CBA-like methods [2, 4, 6, 8, 10]. Unfortunately, support-confidence-based methods show their limits on imbalanced data sets. Indeed, rules with high confidence can also be negatively correlated. [11, 12] propose new methods based on correlation measure to overcome this weakness. However, when

considering a n-class imbalanced context, even a correlation measure is not satisfactory: a rule can be positively correlated with two different classes what leads to conflicting rules. The common problem of these approaches is that they are OVA (one-vs-all) methods, i.e., they split the classification task into n two-class classification tasks (positives vs negatives) and, for each sub-task, look for rules that are relevant in the *positive* class and irrelevant for the union of the other classes. Notice also that the popular emerging patterns (EPs introduced in [13]) and the associated EPs-based classifiers (see e.g. [14] for a survey) are following the same principle. Thus, they can lead to conflicting EPs.

In order to improve state-of-the-art approaches for associative classification when considering multiclass imbalanced training sets, our contribution is twofold. First, we propose an OVE (one-vs-each) method that avoids some of the problems observed with typical CBA-like methods. Indeed, we formally characterize the association rules that can be used for classification purposes when considering that a rule has to be relevant for one class and irrelevant for every other class (instead of being irrelevant for their union). Next, we designed a constrained hill climbing technique that automatically tunes the many parameters (frequency thresholds) that are needed. The paper is organized as follows: Section 2 provides the needed definitions. Section 3 discusses the relevancy of the rules extracted thanks to the algorithm presented in Section 4. Section 5 describes how the needed parameters are automatically tuned. Section 6 provides our experimental study on various benchmark data sets. Section 7 briefly concludes.

2 Definitions

Let $\mathcal{C}$ be the set of classes and n its cardinality. Let $\mathcal{A}$ be the set of Boolean attributes. An object o is defined by the subset of attributes that holds for it, i.e., $o \subseteq \mathcal{A}$. The data in Table 1 illustrate the various definitions. It provides 11 classified objects $(o_k)_{k\in1\ldots11}$. Each of them is described with some of the 6 attributes $(a_l)_{l\in1\ldots6}$ and belongs to one class $(c_i)_{i\in1\ldots3}$. This is a toy labeled transactional data set that can be used to learn an associative classifier that may predict the class value among the three possible ones.

2.1 Class Association Rule

A *Class Association Rule* (CAR) is an ordered pair $(X, c) \in 2^{\mathcal{A}} \times \mathcal{C}$. X is the *body* of the CAR and c its *target class*.

Example 1. In Tab. 1, $(\{a_1, a_5\}, c_3)$ is a CAR. $\{a_1, a_5\}$ is the body of this CAR and c_3 its target class.

2.2 Per-class Frequency

Given a class $d \in \mathcal{C}$ and a set $\mathcal{O}_d$ of objects belonging to this class, the *frequency* of a CAR (X, c) in d is $|\{o \in \mathcal{O}_d | X \subseteq o\}|$. Since the frequency of (X, c) in d

Table 1. Eleven classified objects.

		$\mathcal{A}$						$\mathcal{C}$		
		a_1	a_2	a_3	a_4	a_5	a_6	c_1	c_2	c_3
$\mathcal{O}_{c_1}$	o_1	.		.	.			.		
	o_2	.	.	.	.	.		.		
	o_3	.	.		.			.		
	o_4	.	.	.		.		.		
	o_5	.	.	.		.		.		
$\mathcal{O}_{c_2}$	o_6		.		.		.		.	
	o_7	.	.		.				.	
	o_8	.	.			.			.	
$\mathcal{O}_{c_3}$	o_9	.				.				.
	o_{10}	.				.				.
	o_{11}					.				.

does not depend on c, it is denoted $f_d(X)$. Given $d \in \mathcal{C}$ and a related frequency threshold $\gamma \in \mathbb{N}$, (X, c) is *frequent* (resp. *infrequent*) in d iff $f_d(X)$ is at least (resp. strictly below) γ.

Example 2. In Tab. 1, the CAR $(\{a_1, a_5\}, c_3)$ has a frequency of 3 in c_1, 1 in c_2 and 2 in c_3. Hence, if a frequency threshold $\gamma = 2$ is associated to c_3, $(\{a_1, a_5\}, c_3)$ is frequent in c_3.

With the same notations, the *relative frequency* of a CAR (X, c) in d is $\frac{f_d(X)}{|\mathcal{O}_d|}$.

2.3 Interesting Class Association Rule

Without any loss of generality, consider that $\mathcal{C} = \{c_i | i \in 1 \dots n\}$. Given $i \in 1 \dots n$ and $(\gamma_{i,j})_{j \in 1 \dots n} \in \mathbb{N}^n$ (n per-class frequency thresholds pertaining to each of the n classes), a CAR (X, c_i) is said *interesting* iff:

1. it is frequent in c_i, i.e., $f_{c_i}(X) \geq \gamma_{i,i}$
2. it is infrequent in every other class, i.e., $\forall j \neq i, f_{c_j}(X) < \gamma_{i,j}$
3. any more general CAR is frequent in at least one class different from c_i, i.e., $\forall Y \subset X, \exists j \neq i | f_{c_j}(Y) \geq \gamma_{i,j}$ (*minimal body* constraint).

Example 3. In Tab. 1, assume that the frequency thresholds $\gamma_{3,1} = 4$, $\gamma_{3,2} = 2$, and $\gamma_{3,3} = 2$ are respectively associated to c_1, c_2, and c_3. Although it is frequent in c_3 and infrequent in both c_1 and c_2, $(\{a_1, a_5\}, c_3)$ is not an interesting CAR since $\{a_5\} \subset \{a_1, a_5\}$ and $(\{a_5\}, c_3)$ is neither frequent in c_1 nor in c_2.

3 Relevancy of the Interesting Class Association Rules

3.1 Selecting Better Rules

Constructing a CAR-based classifier means selecting relevant CARs for classification purposes. Hence, the space of CARs is to be split into two: the *relevant*

CARs and the *irrelevant* ones. Furthermore, if $\preceq$ denotes a relevancy (possibly partial) order on the CARs, there should not be a rule r from the relevant CARs and a rule s from the irrelevant CARs s.t. $r \preceq s$. If this never happens, we say that the frontier between relevant and irrelevant CARs is *sound*. Notice that [3] uses the same kind of argument but conserves a one-vs-all perspective.

Using the "Global Frequency + Confidence" Order. The influential work from [2] has been based on a frontier derived from the conjunction of a global frequency (sum of the per-class frequencies for all classes) threshold and a confidence (ratio between the per-class frequency in the target class and the global frequency) threshold. Let us consider the following partial order $\preceq_1$: $\forall (X,Y) \in (2^{\mathcal{A}})^2$, $\forall c \in \mathcal{C}$,

$$(X,c) \preceq_1 (Y,c) \Leftrightarrow \forall d \in \mathcal{C}, \begin{cases} f_c(X) \leq f_d(Y) & \text{if } c = d \\ f_c(X) \geq f_d(Y) & \text{otherwise.} \end{cases}$$

Obviously, $\preceq_1$ is a sensible relevancy order. However, as emphasized in the example below, the frontier drawn by the conjunction of a global frequency threshold and a confidence threshold is not sound w.r.t. $\preceq_1$.

Example 4. Assume a global frequency threshold of 5 and a confidence threshold of $\frac{3}{5}$. In Tab. 1, the CAR $(\{a_3\}, c_1)$ is not (globally) frequent. Thus it is on the irrelevant side of the frontier. At the same time, $(\{a_4\}, c_1)$ is both frequent and with a high enough confidence. It is on the relevant side of the frontier. However, $(\{a_3\}, c_1)$ correctly classifies more objects of $\mathcal{O}_1$ than $(\{a_4\}, c_1)$ and it applies on less objects outside $\mathcal{O}_1$. So $(\{a_4\}, c_1) \preceq_1 (\{a_3\}, c_1)$.

Using the "Emergence" Order. *Emerging patterns* have been introduced in [13]. Here, the frontier between relevancy and irrelevancy relies on a growth rate threshold (ratio between the relative frequency in the target class and the relative frequency in the union of all other classes). As emphasized in the example below, the low number of parameters (one growth rate threshold for each of the n classes) does not support a fine tuning of this frontier.

Example 5. Assume a growth rate threshold of $\frac{8}{5}$. In Tab. 1, the CAR $(\{a_1\}, c_1)$ has a growth rate of $\frac{3}{2}$. Thus it is on the irrelevant side on the frontier. At the same time, $(\{a_2\}, c_1)$ has a growth rate of $\frac{8}{5}$. It is on the relevant side of the frontier. However $(\{a_1\}, c_1)$ correctly classifies more objects of $\mathcal{O}_1$ than $(\{a_2\}, c_1)$ and more clearly differentiates objects in $\mathcal{O}_1$ from those in $\mathcal{O}_2$.

Using the "Interesting" Order. The frontier drawn by the growth rates is sound w.r.t. $\preceq_1$. So is the one related to the so-called interesting CARs. Nevertheless, the latter can be more finely tuned so that the differentiation between two classes is better performed. Indeed, the set of interesting CARs whose target class is c_i is parametrized by n thresholds $(\gamma_{i,j})_{j \in 1 \ldots n}$: one frequency threshold

$\gamma_{i,i}$ and $n-1$ infrequency thresholds for each of the $n-1$ other classes (instead of one for all of them). Hence, to define a set of interesting CARs targeting every class, n^2 parameters enable to finely draw the frontier between relevancy and irrelevancy.

In practice, this quadratic growth of the number of parameters can be seen as a drawback for the experimenter. Indeed, in the classical approaches presented above, this growth is linear and finding the proper parameters already appears as a dark art. This issue will be solved in Section 5 thanks to an automatic tuning of the frequency thresholds.

3.2 Preferring General Class Association Rules

The minimal body constraint avoids redundancy in the set of interesting CARs. Indeed, it can easily be shown that, for every CAR (X, c), frequent in c and infrequent in every other class, it exists a body $Y \subseteq X$ s.t. (Y, c) is interesting and $\forall Z \subset Y, (Z, c)$ is not. Preferring shorter bodies means focusing on more general CARs. Hence, the interesting CARs are more prone to be applicable to new unclassified objects. Notice that the added-value of the minimal body constraint has been well studied in previous approaches for associative classification (see, e.g., [5, 7]).

4 Computing and Using the Interesting Class Association Rules

Let us consider n classes $(c_i)_{i\in 1\ldots n}$ and let us assume that Γ denotes a $n \times n$ matrix of frequency thresholds. The i^{th} line of Γ pertains to the subset of interesting CARs whose target class is c_i. The j^{th} column of Γ pertains to the frequency thresholds in c_j. Given Γ and a set of classified objects, we discuss how to efficiently compute the complete set of interesting CARs.

4.1 Enumeration

The complete extraction of the interesting CARs is performed one target class after another. Given a class $c_i \in \mathcal{C}$, the enumeration strategy of the candidate CARs targetting c_i is critical for performance issues. The search space of the CAR bodies, partially ordered by $\subseteq$, has a lattice structure. It is traversed in a breadth-first way. The two following properties enable to explore only a small part of it without missing any interesting CAR:

1. If (Y, c_i) is not frequent in c_i, neither is any (X, c_i) with $Y \subseteq X$.
2. If (Y, c_i) is an interesting CAR, any (X, c_i) with $Y \subset X$ does not have a minimal body.

Such CAR bodies Y are collected into a prefix tree. When constructing the next level of the lattice, every CAR body in the current level is enlarged s.t. it does not become a superset of a body in the prefix tree. In this way, entire sublattices, which cannot contain bodies of interesting CARs, are ignored.

4.2 Algorithm

Algorithm 1 details how the extraction is performed. `parents` denotes the current level of the lattice (i.e., a list of CAR bodies). `futureParents` is the next level. `forbiddenPrefixes` is the prefix tree of forbidden subsets from which is computed `forbiddenAtts`, the list of attributes that are not allowed to enlarge `parent` (a CAR body in the current level) to give birth to its `child`ren (bodies in the next level).

```
forbiddenPrefixes ← ∅
parents ← [∅]
while parents ≠ [] do
    futureParents ← ∅
    for all parent ∈ parents do
        forbiddenAtts ← forbiddenAtts(forbiddenPrefixes, parent)
        for all attribute > lastAttribute(parent) do
            if attribute ∉ forbiddenAtts then
                child ← constructChild(parent, attribute)
                if f_{c_i}(child) ≥ γ_{i,i} then
                    if interesting(child) then
                        output (child, c_i)
                        insert(child, forbiddenPrefixes)
                    else
                        futureParents ← futureParents ∪ {child}
                else
                    insert(child, forbiddenPrefixes)
        parents ← parents \ {parent}
    parents ← futureParents
```

Algorithm 1: EXTRACT(c_i: target class)

4.3 Simultaneously Enforcing Frequency Thresholds in All Classes

Notice that, along the extraction of the interesting CARs targeting c_i, all frequency thresholds $(\gamma_{i,j})_{j\in 1\ldots n}$ are simultaneously enforced. To do so, every CAR body in the lattice is bound to n bitsets related to the $(\mathcal{O}_{c_i})_{i\in 1\ldots n}$. Thus, every bit stands for the match ('1') or the mismatch ('0') of an object and bitwise ANDs enables an incremental and efficient computation of `child`ren's bitsets.

Alternatively, the interesting CARs targeting c_i could be obtained by computing the $n-1$ sets of emerging patterns between c_i and every other class c_j (with $\frac{\gamma_{i,i}|\mathcal{O}_j|}{\gamma_{i,j}|\mathcal{O}_i|}$ as a growth rate), one by one, and intersecting them. However, the time complexity of n^2-n extractions of loosely constrained CARs is far worse than ours (n extractions of tightly constrained CARs). When n is large, it prevents from automatically tuning the parameters with a hill climbing technique.

4.4 Classification

When an interesting CAR is output, we can output its vector of relative frequencies in all classes at no computational cost. Then, for a given unclassified object $o \subseteq \mathcal{A}$, its likeliness to be in the class c_i is quantifiable by $l(o, c_i)$ which is the sum of the relative frequencies in c_i of all interesting CARs applicable to o:

$$l(o, c_i) = \sum_{c \in \mathcal{C}} \left(\sum_{\text{interesting } (X,c) \text{ s.t. } X \subseteq o} \left(\frac{f_{c_i}(X)}{|\mathcal{O}_{c_i}|} \right) \right)$$

Notice that the target class of an interesting CAR does not hide the exceptions it may have in the other classes. The class $c_{\max}$ related to the greatest likeliness value $l(o, c_{\max})$ is where to classify o. $\min_{i \neq \max}(\frac{l(o,c_{\max})}{l(o,c_i)})$ quantifies the certainty of the classification of o in the class $c_{\max}$ rather than c_i (the other class with which the confusion is the greatest). This "certainty measure" may be very valuable in cost-sensitive applications.

5 Automatic Parameter Tuning

It is often considered that manually tuning the parameters of an associative classification method, like our CAR-based algorithm, borders the dark arts. Indeed, our algorithm from Sec. 4 requires a n-by-n matrix Γ of input parameters. Fortunately, analyzing the way the interesting CARs apply to the learning set, directly indicates what frequency threshold in Γ should be modified to probably improve the classification. We now describe how to algorithmically tune Γ to obtain a set of interesting CARs that is well adapted to classification purposes. Due to space limitations, the pseudo-code of this algorithm, called `fitcare`[3], is only available in an associated technical report [15].

5.1 Hill Climbing

The `fitcare` algorithm tunes Γ following a hill climbing strategy.

Maximizing the Minimal Global Growth Rate. Section 4.4 mentioned the advantages of not restricting the output of a CAR to its target class (its frequencies in every class are valuable as well). With the same argument applied to the global set of CARs, the hill climbing technique, embedded within `fitcare`, maximizes *global growth rates* instead of other measures (e.g., the number of correctly classified objects) where the loss of information is greater.

Given two classes $(c_i, c_j) \in \mathcal{C}^2$ s.t. $i \neq j$, the *global growth rate* $g(c_i, c_j)$ quantifies, when classifying the objects from $\mathcal{O}_{c_i}$, the confusion with the class c_j. The greater it is, the less confusion made. We define it as follows:

$$g(c_i, c_j) = \frac{\sum_{o \in \mathcal{O}_{c_i}} l(o, c_i)}{\sum_{o \in \mathcal{O}_{c_i}} l(o, c_j)}$$

[3] `fitcare` is the recursive acronym for `f`itcare `i`s `t`he `c`lass `a`ssociation `r`ule `e`xtractor.

From a set of interesting CARs, `fitcare` computes all $n^2 - n$ global growth rates. The maximization of the minimal global growth rate drives the hill climbing, i.e., `fitcare` tunes Γ so that this rate increases. When no improvement can be achieved on the smallest global growth rate, `fitcare` attempts to increase the second smallest (while not decreasing the smallest), etc. `fitcare` terminates when a maximum is reached.

Choosing one $\gamma_{i,j}$ to lower. Instead of a random initialization of the parameters (a common practice in hill climbing techniques), Γ is initialized with high frequency thresholds. The hill climbing procedure only lowers these parameters, one at a time, and by decrements of 1. However, we will see, in Sec. 5.2, that such a modification leads to lowering other frequency thresholds if Γ enters an undesirable state.

The choice of the parameter $\gamma_{i,j}$ to lower depends on the global growth rate $g(c_i, c_j)$ to increase. Indeed, when classifying the objects from $\mathcal{O}_{c_i}$, different causes lead to a confusion with c_j. To discern the primary cause, every class at the denominator of $g(c_i, c_j)$ is evaluated separately:

$$\begin{pmatrix} \sum_{o \in \mathcal{O}_{c_i}} \sum_{\text{interesting } (X,c_1) \text{ s.t. } X \subseteq o} \left(\frac{f_{c_j}(X)}{|\mathcal{O}_{c_j}|} \right) \\ \sum_{o \in \mathcal{O}_{c_i}} \sum_{\text{interesting } (X,c_2) \text{ s.t. } X \subseteq o} \left(\frac{f_{c_j}(X)}{|\mathcal{O}_{c_j}|} \right) \\ \vdots \\ \sum_{o \in \mathcal{O}_{c_i}} \sum_{\text{interesting } (X,c_n) \text{ s.t. } X \subseteq o} \left(\frac{f_{c_j}(X)}{|\mathcal{O}_{c_j}|} \right) \end{pmatrix}$$

The greatest term is taken as the primary cause for $g(c_i, c_j)$ to be small. Usually it is either the i^{th} term (the interesting CARs targeting c_i are too frequent in c_j) or the j^{th} one (the interesting CARs targeting c_j are too frequent in c_i). This term directly indicates what frequency threshold in Γ should be preferably lowered. Thus, if the i^{th} (resp. j^{th}) term is the greatest, $\gamma_{i,j}$ (resp. $\gamma_{j,i}$) is lowered. Once Γ modified and the new interesting CARs extracted, if $g(c_i, c_j)$ increased, the new Γ is committed. If not, Γ is rolled-back to its previous value and the second most promising $\gamma_{i,j}$ is decremented, etc.

5.2 Avoiding Undesirable Parts of the Parameter Space

Some values for Γ are obviously bad. Furthermore, the hill climbing technique cannot properly work if too few or too many CARs are interesting. Hence, `fitcare` avoids these parts of the parameter space.

Sensible Constraints on Γ. The relative frequency of an interesting CAR targeting c_i should obviously be strictly greater in c_i than in any other class:

$$\forall i \in 1 \ldots n, \forall j \neq i, \frac{\gamma_{i,j}}{|\mathcal{O}_{c_j}|} < \frac{\gamma_{i,i}}{|\mathcal{O}_{c_i}|}$$

Furthermore, the set of interesting CARs should be *conflictless*, i.e., if it contains (X, c_i), it must not contain (Y, c_j) if $Y \subseteq X$. Thus, an interesting CAR targeting c_i must be strictly more frequent in c_i than any interesting CAR whose target class is not c_i:

$$\forall i \in 1 \dots n, \forall j \neq i, \gamma_{i,j} < \gamma_{j,j}$$

Whenever a modification of Γ violates one of these two constraints, every $\gamma_{i,j}$ $(i \neq j)$ in cause is lowered s.t. Γ reaches another sensible state. Then, the extraction of the interesting CARs is performed.

Minimal Positive Cover Rate Constraint. Given a class $c \in \mathcal{C}$, the *positive cover rate* of c is the proportion of objects in $\mathcal{O}_c$ that are covered by at least one interesting CAR targeting c, i.e., $\frac{|\{o \in \mathcal{O}_c | \exists \text{ interesting } (X,c) \text{ s.t. } X \subseteq o\}|}{|\mathcal{O}_c|}$. Obviously, the smaller the positive cover rate of c, the worse the classification in c.

By default, `fitcare` forces the positive cover rates of every class to be 1 (every object is positively covered). Thus, whenever interesting CARs, with c_i as a target class, are extracted, the positive cover rate of c_i is returned. If it is not 1, $\gamma_{i,i}$ is lowered by 1 and the interesting CARs are extracted again.

Notice that `fitcare` lowers $\gamma_{i,i}$ until $\mathcal{O}_{c_i}$ is entirely covered but not more. Indeed, this could bring a disequilibrium between the average number of interesting CARs applying to the objects in the different classes. If this average in $\mathcal{O}_{c_i}$ is much higher than that of $\mathcal{O}_{c_j}$, $g(c_i, c_j)$ would be artificially high and $g(c_j, c_i)$ artificially low. Hence, the hill climbing strategy would be biased.

On some difficult data sets (e.g., containing misclassified objects), it may be impossible to entirely cover some class c_i while verifying $\forall i \in 1 \dots n, \forall j \neq i, \frac{\gamma_{i,j}}{|\mathcal{O}_{c_j}|} < \frac{\gamma_{i,i}}{|\mathcal{O}_{c_i}|}$. That is why, while initializing Γ, a looser minimal positive cover rate constraint may be decided.

Here is how the frequency thresholds in the i^{th} line of Γ are initialized (every line being independently initialized):

$$\forall j \in 1 \dots n, \gamma_{i,j} = \begin{cases} |\mathcal{O}_{c_j}| \text{ if } i = j \\ |\mathcal{O}_{c_j}| - 1 \text{ otherwise} \end{cases}$$

The interesting CARs targeting c_i are collected with $\textsc{Extract}(c_i)$. Most of the time, the frequency constraint in c_i is too high for the interesting CARs to entirely cover $\mathcal{O}_{c_i}$. Hence `fitcare` lowers $\gamma_{i,i}$ (and the $(\gamma_{i,j})_{j \in 1 \dots n}$ s.t. $\forall i \in 1 \dots n, \forall j \neq i, \frac{\gamma_{i,j}}{|\mathcal{O}_{c_j}|} < \frac{\gamma_{i,i}}{|\mathcal{O}_{c_i}|}$) until $\mathcal{O}_{c_i}$ is entirely covered. If $\gamma_{i,i}$ reaches 0 but the positive cover rate of c_i never was 1, the minimal positive cover rate constraint is loosened to the greatest rate encountered so far. The frequency thresholds related to this greatest rate constitute the i^{th} line of Γ when the hill climbing procedure starts.

6 Experimental Results

The `fitcare` algorithm has been implemented in `C++`. We performed an empirical validation of its added-value on various benchmark data sets. The `LUCS-KDD` software library [16] provided the discretized versions of the UCI data sets [17] and a `Java` implementation of `CPAR`. Notice that we name the data sets according to Coenen's notation, e.g., the data set "breast.D20.N699.C2" gathers 699 objects described by 20 Boolean attributes and organized in 2 classes. To put the focus on imbalanced data sets, the repartition of the objects into the classes is mentioned as well. Bold faced numbers of objects indicate minor classes, i.e., classes having, at most, half the cardinality of the largest class.

The global and the per-class accuracies of `fitcare` are compared to that of `CPAR`, one of the best CBA-like methods designed so far. The results, reported in Tab. 2, were obtained after 10-fold stratified cross validations.

Table 2. Experimental results of `fitcare` and comparison with `CPAR`.

Data Sets	*Global*		*Per-class (True Positive rates)*	
	`fitcare`	`CPAR`	`fitcare`	`CPAR`
anneal.D73.N898.C6 **8**/**99**/684/**0**/**67**/**40**	92.09	**94.99**	**87.5**/46.46/98.09/-/**100**/90	17/**90.24**/99.44/-/**100**/**96.25**
breast.D20.N699.C2 458/241	82.11	**92.95**	73.36/98.75	98.58/84.68
car.D25.N1728.C4 1210/**384**/**69**/**65**	**91.03**	80.79	98.67/**73.43**/**66.66**/**78.46**	92.25/58.74/46.03/23.67
congres.D34.N435.C2 267/168	88.96	**95.19**	89.13/88.69	97.36/92.31
cylBands.D124.N540.C2 228/312	**68.7**	68.33	30.7/92.94	61.99/79.99
dermatology.D49.N366.C6 72/112/61/**52**/**49**/**20**	77.86	**80.8**	80.55/82.14/62.29/**78.84**/79.59/**85**	80.65/88.86/67.71/77.94/**96.67**/46
glass.D48.N214.C7 70/76/**17**/**0**/**13**/**9**/**29**	**72.89**	64.1	80/68.42/**23.52**/-/**76.92**/**100**/86.2	54.49/65.71/0/-/45/30/**90**
heart.D52.N303.C5 164/**55**/**36**/**35**/**13**	**55.44**	55.03	81.7/**21.81**/19.44/**34.28**/**23.07**	78.68/14.86/**23.26**/23.79/10
hepatitis.D56.N155.C2 **32**/123	**85.16**	74.34	**50**/95.93	45.05/94.37
horsecolic.D85.N368.C2 232/136	81.25	**81.57**	81.46/80.88	85.69/76.74
iris.D19.N150.C3 50/50/50	**95.33**	**95.33**	100/94/92	100/91.57/96.57
nursery.D32.N12960.C5 4320/**2**/**328**/4266/4044	**98.07**	78.59	100/-/**81.7**/96.78/99.45	77.64/-/21.24/73.53/98.74
pima.D42.N768.C2 500/268	72.78	**75.65**	84.2/51.49	78.52/69.03
ticTacToe.D29.N958.C2 626/332	65.76	**71.43**	63.73/69.57	76.33/63
waveform.D101.N5000.C3 1657/1647/1696	**77.94**	70.66	59.56/88.4/85.73	72.87/69.13/71.67
wine.D68.N178.C3 59/71/48	**95.5**	88.03	96.61/94.36/95.83	85.38/87.26/94.67
Arithmetic Means	**81.3**	79.24	**76.57**	69.08

6.1 2-Class vs Multiclass Problem

2-class Problem. Five of the seven data sets where `CPAR` outperforms `fitcare` correspond to well-balanced 2-class problems, where the minimal positive cover constraint has to be loosened for one of the classes (see Sec. 5.2). On the two remaining 2-class data sets, which do not raise this issue (cylBands and hepatitis), `fitcare` has a better accuracy than `CPAR`.

Multiclass Problem. `fitcare` significantly outperforms `CPAR` on all the nine multiclass data sets but two – anneal and dermatology – on which `fitcare` lies slightly behind `CPAR`. On the nursery data, the improvement in terms of global accuracy even reaches 25% w.r.t. `CPAR`.

6.2 True Positive Rates in Minor Classes

When considering imbalanced data sets, True Positive rates (TPr) are known to better evaluate classification performances. When focusing on the TPr in the minor classes, `fitcare` clearly outperforms `CPAR` in 14 minor classes out of the 20 (bold values). Observe also that the 2-class data sets with a partial positive cover of the largest class have a poor global accuracy but the TPr of the smallest classes often are greater than `CPAR`'s (see breast, horsecolic, ticTacToe).

Compared to `CPAR`, `fitcare` presents better arithmetic means in both the global and the per-class accuracies. However, the difference is much greater with the latter measure. Indeed, as detailed in Sec. 5.1, `fitcare` is driven by the minimization of the confusion between every pair of classes (whatever their sizes). As a consequence, `fitcare` optimizes the True Positive rates. In the opposite, `CPAR` (and all one-vs-all approaches), focusing only on the global accuracy, tends to over-classify in the major classes.

7 Conclusion

Association rules have been extensively studied along the past decade. The CBA proposal has been the first associative classification technique based on a "support-confidence" ranking criterion [2]. Since then, many other CBA-like approaches have been designed. Even if suitable for typical two-class problems, it appears that support and confidence constraints are inadequate for selecting rules in multiclass imbalanced training data sets. Other approaches (see, e.g., [11, 14, 12]) address the problem of imbalanced data sets but show their limits when considering more than 2 classes. We analyzed the limits of all these approaches, suggesting that a common weakness relies on their one-vs-all principle. We proposed a solution to these problems: our associative classification method extracts the so-called interesting class association rules w.r.t. a one-vs-each principle. It computes class association rules that are frequent in the positive class and infrequent in every other class taken separately (instead of their union). Tuning the large number of parameters required by this approach may appear as a bottleneck. Therefore, we designed an automatic tuning method that relies on a hill-climbing strategy. Empirical results have confirmed that our proposal is quite promising for multiclass imbalanced data sets.

Acknowledgments. This work is partly funded by EU contract IST-FET IQ FP6-516169 and by the French contract ANR ANR-07-MDCO-014 Bingo2. We would like to thank an anonymous reviewer for its useful concerns regarding the relevancy of our approach. Unfortunately, because of space restrictions, we could not address them all in this article.

References

1. Agrawal, R., Imielinski, T., Swami, A.N.: Mining Association Rules Between Sets of Items in Large Databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, ACM Press (1993) 207–216
2. Liu, B., Hsu, W., Ma, Y.: Integrating Classification and Association Rule Mining. In: Proceedings of the Fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, AAAI Press (1998) 80–86
3. Bayardo, R., Agrawal, R.: Mining the Most Interesting Rules. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press (1999) 145–154
4. Li, W., Han, J., Pei, J.: CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules. In: Proceedings of the First IEEE International Conference on Data Mining, IEEE Computer Society (2001) 369–376
5. Boulicaut, J.F., Crémilleux, B.: Simplest Rules Characterizing Classes Generated by Delta-free Sets. In: Proceedings of the Twenty-Second Annual International Conference Knowledge Based Systems and Applied Artificial Intelligence, Springer (2002) 33–46
6. Yin, X., Han, J.: CPAR: Classification Based on Predictive Association Rules. In: Proceedings of the Third SIAM International Conference on Data Mining, SIAM (2003) 369–376
7. Baralis, E., Chiusano, S.: Essential Classification Rule Sets. ACM Transactions on Database Systems **29**(4) (2004) 635–674
8. Bouzouita, I., Elloumi, S., Yahia, S.B.: GARC: A New Associative Classification Approach. In: Proceedings of the Eight International Conference on Data Warehousing and Knowledge Discovery, Springer (2006) 554–565
9. Freitas, A.A.: Understanding the Crucial Differences Between Classification and Discovery of Association Rules – A Position Paper. SIGKDD Explorations **2**(1) (2000) 65–69
10. Wang, J., Karypis, G.: HARMONY: Efficiently Mining the Best Rules for Classification. In: Proceedings of the Fifth SIAM International Conference on Data Mining, SIAM (2005) 34–43
11. Arunasalam, B., Chawla, S.: CCCS: A Top-down Associative Classifier for Imbalanced Class Distribution. In: Proceedings of the Twelveth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press (2006) 517–522
12. Verhein, F., Chawla, S.: Using Significant Positively Associated and Relatively Class Correlated Rules for Associative Classification of Imbalanced Datasets. In: Proceedings of the Seventh IEEE International Conference on Data Mining, IEEE Computer Society 679–684
13. Dong, G., Li, J.: Efficient Mining of Emerging Patterns: Discovering Trends and Differences. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press (1999) 43–52
14. Ramamohanarao, K., Fan, H.: Patterns Based Classifiers. World Wide Web **10**(1) (2007) 71–83
15. Cerf, L., Gay, D., Selmaoui, N., Boulicaut, J.F.: Technical Notes on `fitcare`'s Implementation. Technical report, LIRIS (april 2008)
16. Coenen, F.: The LUCS-KDD software library (2004) `http://www.csc.liv.ac.uk/~frans/KDD/Software/`.
17. Newman, D., Hettich, S., Blake, C., Merz, C.: UCI Repository of Machine Learning Databases (1998) `http://www.ics.uci.edu/~mlearn/MLRepository.html`.