Under consideration for publication in Knowledge and Information Systems

From Global to Local and Viceversa: Uses of Associative Rule Learning for Classification in Imprecise Environments

Gianni Costa, Giuseppe Manco, Riccardo Ortale and Ettore Ritacco

ICAR-CNR, Via Bucci 41c, 87036 Rende (CS), Italy

Abstract. We propose two models for improving the performance of rule-based classification under unbalanced and highly imprecise domains. Both models are probabilistic frameworks aimed to boost the performance of basic rule-based classifiers. The first model implements a *global-to-local* scheme, where the response of a global rule-based classifier is refined by performing a probabilistic analysis of the coverage of its rules. In particular, the coverage of the individual rules is used to learn local probabilistic models, that ultimately refine the predictions from the corresponding rules of the global classifier. The second model implements a dual *local-to-global* strategy, in which single classification rules are combined within an exponential probabilistic model in order to boost the overall performance as a side effect of mutual influence. Several variants of the basic ideas are studied and their performances are thoroughly evaluated and compared with state-of-the-art algorithms on standard benchmark datasets.

Keywords: Rule Learning, Associative Classification, Rarity, Maximum Entropy

1. Introduction

Classification is one of the most extensively studied tasks in machine learning, pattern recognition and data mining. Given a collection of labeled training data, the aim is to learn a suitable model, referred to as a classifier, wherein the regularities in the labeled data are exploited to induce a reasonable approximation (i.e. a hypothesis) on the actual mappings between any data case from the same domain and one of multiple predefined class labels. A classifier is, hence, useful to predict the unknown class of a previously unseen case, on the basis of the other observable features of the same case. Various types of classifiers have been proposed in the literature, that meet several different requirements in a wealth of distinct applicative settings, such as decision trees, rule-based

Received Nov 22, 2010
Revised Jun 24, 2011
Accepted Sep 24, 2011

classifiers, neural networks, naïve Bayes classifiers, support vector machines and statistical classifiers (Duda et al, 2001). In particular, rule learning is a method for inducing minimal rule-based concept descriptions, that can be used for classification. Rule-based classifiers are a mainstay of research in machine learning, because of various desirable properties such as, e.g., their expressiveness and intelligibility to humans as well as their efficiency and effectiveness in classification. Such classifiers have been empirically shown to be effective in processing (sparse) high-dimensional training data with categorical attributes (Wang and Karypis, 2005) and are comparable in performance with other classification methods in several applicative domains (Mitchell, 1997). Unfortunately, like most classification models, rule-based classifiers exhibit a poor classification performance in imprecise (multi-class) learning environments, which are challenging domains wherein cases and classes of primary interest for the learning task are rare. Besides, minority and majority classes can be hardly separable and the cost of misclassifying a case of a minority class as belonging to a predominant class is much higher than the cost of the dual error. Also, training data may be corrupted by noise, which further obstacles the identification of rarities. Imprecise domains are often encountered in practical applications. Examples include fraud detection (Fawcett and Provost, 1997; Phua et al, 2004), intrusion detection (Tang et al, 2007), manufacturing line monitoring (Riddle et al, 1994), risk management, telecommunications management (Ezawa et al, 1996), medical diagnosis (Chawla et al, 2002), text classification (Weiss, 2004) and oil-spill detection in satellite images (Kubat et al, 1998). The peculiarities of such settings pose several challenging issues to traditional algorithms for learning rule-based classifiers, that essentially make the resulting models low sensitive to rarities.

Rarity is clearly the major obstacle. Rare classes corresponds to the well known *class imbalance* issue (Japkowicz, 2000; Japkowicz and Stephen, 2002), i.e. an evenly distribution of classes, such that majority classes overwhelm minority ones. Instead, rare cases are very small portions of the training data, that can be viewed as exceptional sub-concepts seldom occurring within predominant or rare classes. As it is pointed out in (Weiss, 2004), rarity actually prevents conventional algorithms for rule induction from finding and reliably generalizing the regularities within infrequent classes and exceptional cases.

Indeed, class imbalance generally leads to classification models tending to exhibit a high specificity (i.e. capability at recognizing majority classes), coupled with a low sensitivity (i.e. capability at recognizing minority classes).

Rare cases, instead, tend to materialize within the learnt classification models as *small disjuncts* (Holte et al, 1989), i.e. rules covering very few training cases (Weiss, 2000). Small disjuncts were empirically shown to be a major cause of poor predictive performance (Weiss and Hirsh, 2000) and cannot be easily removed without adversely affecting the remaining classification rules.

The foregoing effects of rarity on rule learning are exacerbated by noise. On one hand, the latter may further skew class imbalance. On the other hand, it may also appear to the learner as nearly indistinguishable from rare cases.

Besides rarity and noise, different misclassification costs as well as low class separability also have a role in making conventional rule learning schemes inadequate within imprecise domains.

Cost-sensitive methods (Elkan, 2001; Pazzani et al, 1994) may be used to account for different misclassification costs by explicitly assigning an appropriately higher value to the recognition of minority classes with respect to the identification of majority classes. The overall learning process would thus be biased towards rare classes and the corresponding decision regions within the resulting classification models would have broader boundaries, suitably extended to cover more minority classes, even if at the

expense of an increased number of (misclassified) majority classes. Nonetheless, the domain-specific information on misclassification costs is either seldom known or hardly quantifiable in an objective manner whenever related to domain experts' subjectiveness.

In the last decade, classification based on frequent patterns, also known as *associative classification*, has emerged as a powerful enhancement of conventional rule learning, based on converging research efforts in machine learning and data mining (for instance see (Hämäläinen, 2010)). Precisely, the basic intuition behind associative classification is to substitute conventional rule induction with an association-rule mining step. The resulting classification models, said associative classifiers, consist of class association rules, i.e. suitable association rules meeting some specific constraints. The antecedents of these rules are co-occurrent attribute values, that frequently appear across the training data, while their consequents are suitable values of the target class attribute. Associative classification is in principle better suited for unsupervised predictive modeling within imprecise learning settings: it retains the advantages of traditional rule learning and also tends to achieve a better performance for several reasons. Foremost, while rule induction dilutes rarity and produces overly biased rules, associative classification yields rules with an appropriate degree of generality/specificity, that summarize the whole training data. Also, the individual class association rules catch strong, i.e. frequently occurring, associations between (combinations of) data items and class labels. This is a robust mechanism with which to handle noise in data. Additionally, such associations reflect the inherent semantics of the training data and, thus, have a high discriminative power. The resulting associative classifiers are statistically significant and are hence deemed to properly generalize on unseen data (Cheng et al, 2007). Furthermore, frequent patterns represent a more expressive feature space, where the original training data is likelier to be linearly separable.

One limitation for associative classification, that is particularly relevant in imprecise learning settings, is borrowed from traditional rule learning. More specifically, the decision regions induced by a rule-based classifier and the true distribution of the classes in the space of data do not match. Indeed, classes form regions with irregular and interleaved shapes, whereas the induced decision regions are neatly separated by boundaries parallel to the features of the data space. As a consequence, those cases falling within and close to the boundary of a decision region may be misleadingly predicted as belonging to the class associated with that decision region, even if the true class membership in the surroundings of the boundary is different. This is problematic in imprecise applicative domains, wherein the separability between classes is low, since these form true overlapping (or embedded) regions. In such cases, indeed, the true regions formed by rare classes may be (partly or completely) overlapped by the decision regions associated to the predominant classes and, thus, the recognition of previously unseen cases of the rare classes becomes a major concern.

In this manuscript, we explore associative classification as an enabling mechanism for designing new classification schemes, capable to induce predictive models that effectively discriminate rare classes within imprecise (multi-class) learning environments. We do not deliberately deal with rare cases. As it is pointed out in (Weiss, 2004), the effectiveness of a classification strategy on rare cases cannot be directly evaluated, since these are usually unknown. Notwithstanding, both rare classes and rare cases are argued to be two strongly related facets of rarity, whose issues can be addressed with the same methods. Hence, it is reasonable to expect that if an approach is effective with rare classes, it is also useful for dealing with rare cases.

Two approaches are proposed that look at associative classification from two dual perspectives.

From the *global-to-local* point of view, associative rule learning yields a global

(high level) classification model, whose class assignments are then refined locally to the individual classifier rules. In this regard, one approach essentially builds a hierarchical classification framework, that combines associative rule learning and probabilistic smoothing (Costa et al, 2009). The underlying idea is to use the individual rules of an associative classifier to divide the original training data into as many segments, wherein it is likely that some globally rare cases/classes become less rare. The resulting segments are then used to build as many local probabilistic generative models, that better catch the forms of rarity local to their segments. These probabilistic generative models are then used to refine the predictions from the classifier rules. Two distinct schemes are proposed for tightly integrating associative classification and probabilistic smoothing, that decide the class of an unlabeled case by considering multiple class association rules as well as their corresponding probabilistic generative models.

From the *local-to-global* point of view, instead, associative rule learning provides local data features, that determine global assignments of class probabilities. Therein, in the second approach, the individual rules of an associative classifier are used as features. Given a data case, classification takes into account the predictions from all those rules that are local to the case (i.e. that cover the case). The relevance of a rule with respect to its targeted class determines the weight of the corresponding feature on the discrimination of that particular class. This enables the recognition of minority classes via those classification rules, that are highly representative of such classes (i.e. whose antecedent reflects item co-occurrences that are inherently characteristic of such classes). The maximum entropy framework is used to elegantly and seamlessly integrate associative classification with discriminative learning.

An intensive empirical evaluation shows that both the *global-to-local* and the *local-to-global* approaches are competitive and often superior in accuracy and precision w.r.t. established competitors, while overcoming them in the ability to deal with rare classes.

To summarize, our contributions include:

- A study on associative classification as the basic building block for designing new classification schemes, capable to discriminate rare classes within imprecise (multi-class) domains. The study considers associative classification both from the *global-to-local* and the *local-to-global* perspectives.
- A hierarchical framework for *global-to-local* classification, wherein we exploit probabilistic smoothing to locally refine the classification decisions from a (global) associative classifier.
- A probabilistic framework for *local-to-global* classification, based on maximum entropy modeling, that learns a discriminative classifier, i.e. the conditional probability distribution of classes given an unlabeled case, in which the rules of an associative classifier are viewed as features (local to the case) influencing (global) class probability assignments.
- A thorough comparative experimentation of both the *global-to-local* and the *local-to-global* approaches.

The outline of the rest of this manuscript is as follows. Section 2 introduces notation and preliminaries. Sections 3 and 4 discuss, respectively, the *global-to-local* and *local-to-global* approaches. Section 5 presents the empirical evaluation of both approaches. Section 6 overviews some seminal works from the current literature that are closely related to our approaches. Finally, section 7 concludes and highlights promising directions of further research.

2. Preliminaries

We begin by introducing the notation used throughout the manuscript and some basic notions. Let $\mathcal{D}$ be a relation storing the labeled training cases. Also, let the schema of $\mathcal{D}$ be a set $\mathcal{A} = \{A_1 : Dom(A_1), \ldots, A_n : Dom(A_n), L : \mathcal{L}\}$ of descriptive attributes. In particular, features $A_1, \ldots, A_n$ are defined over as many categorical or numeric domains, whereas the target class attribute L is a categorical feature. The generic labeled training case $t \in \mathcal{D}$ is a structured tuple, i.e. $t \in Dom(A_1) \times \ldots Dom(A_n) \times Dom(L)$. Each tuple t can also be equivalently represented in a transactional form. Therein, assume that $\mathcal{M} = \{i_1, \ldots, i_m\}$ is a finite set of items denoting relationships between any attribute of $\mathcal{A}$ but L and a corresponding value. Precisely, the generic item i has the form $A = v$ where $A \in \mathcal{A} - L$. In our formulation, $v \in Dom(A)$ if A is a categorical attribute. Otherwise, if A is a numeric attribute, v stands for the label of some suitable range of numeric values, whose center is closest in Euclidean distance to the original value of A (more details on the discretization of numeric attributes are provided within section 5).

Any unlabeled case I defined over $\mathcal{A}$ can be represented as some suitable subset of items in $\mathcal{M}$. Notice that there must be exactly one item in I for each attribute of the relational schema $\mathcal{A}$. This is concisely expressed by means of the $\subset$ operator, whose meaning is revised as follows $I \subset \mathcal{M} = \{i_{j_1}, \ldots, i_{j_n} | \, i_{j_h} \in \mathcal{M} \wedge attr(i_{j_h}) = A_h, \, \forall h = 1, \ldots, n\}$, where notation $attr(\cdot)$ indicates the attribute referred to by the individual items of I. Viewed from this perspective, a labeled case over $\mathcal{A}$ simply becomes an unlabeled case supplemented with its corresponding class label. Let $\mathcal{L}$ be a finite domain of class labels, the original dataset $\mathcal{D}$ can thus be equivalently expressed in transactional form over $\mathcal{M}$ as a collection $\mathcal{D} = \{t_1, \ldots, t_n\}$, in which the generic labeled case is represented as $t = I \cup \{class(t)\}$, where $I \subset \mathcal{M}$ and $class(t) \in \mathcal{L}$ denotes the class label of t. Henceforth, for convenience, we shall refer to the transactional representation and use the corresponding notation.

A number of definitions recalled throughout the manuscript are reported next.

Definition 2.1 (Class association rule). According to the original proposal in (Liu et al, 1998), a class association rule (CAR) $r : I \rightarrow c$ is an implicative pattern, that catches the association (i.e. the co-occurrence) in $\mathcal{D}$ of some subset of items $I \subset \mathcal{M}$ with a class label c from $\mathcal{L}$. □

The notions of support, coverage and confidence are typically employed to define the interestingness of a rule r.

Definition 2.2 (Support of a class association rule). Let $\mathcal{D}$ be a set of training cases. A training case $t \in \mathcal{D}$ is said to *support* rule $r : I \rightarrow c$ if it holds that $(I \cup \{c\}) \subseteq t$. The support count of r, denoted by $\sigma(r)$, is the overall number of training cases that support r, i.e., $\sigma(r) = |\{t \in \mathcal{D} | (I \cup \{c\}) \subseteq t\}|$. The support of r is instead the fraction of training cases supporting r, i.e., $supp(r) = \frac{\sigma(r)}{|\mathcal{D}|}$, where $|\mathcal{D}|$ indicates the cardinality of $\mathcal{D}$. □

Support is useful to avoid spurious rules. Intuitively, rule antecedents with high support in the individual classes capture the inherent semantics of the underlying data, rather than being artifacts.

Definition 2.3 (Coverage of a class association rule). Let $\mathcal{D}$ be a set of training cases. Rule $r : I \rightarrow c$ is said to *cover* a training case $t \in \mathcal{D}$ (and, dually, t is said to trigger or fire r) if the condition $I \subseteq (t - \{class(t)\})$ holds. The set of training cases covered by r is denoted by $\mathcal{D}_r = \{t \in \mathcal{D} | I \subseteq (t - \{class(t)\})\}$. Hence, the coverage of r can

be defined as the fraction of cases in $\mathcal{D}$ that are covered by r, i.e. $coverage(r) = \frac{|\mathcal{D}_r|}{|\mathcal{D}|}$. Analogously, the foresaid rule $r : I \rightarrow c$ is said to *cover* an unlabeled training case I' if it holds that $I \subseteq I'$. □

Definition 2.4 (Confidence of a class association rule). The confidence of a rule r, denoted by $conf(r)$, is the ratio of support to coverage, i.e. $conf(r) = \frac{supp(r)}{coverage(r)}$. □

Confidence measures the predictive strength of a CAR.

Although the traditional support and confidence framework allows to effectively discover all the required class association rules, it still produces uninteresting rules when the class distribution is imbalanced. The point is that, in such cases, confidence is not a reliable measure of the interestingness of a rule (Tatti, 2008), since it does not properly take into account the actual implicative strength of the rule, whose antecedent and consequent can be negatively correlated (Arunasalam and Chawla, 2006; Antonie and Zaïane, 2004). To overcome such a limitation, we also consider the degree of positive correlation between the antecedent and the consequent of a rule.

Definition 2.5 (CAR correlation). The correlation of a rule $r : I \rightarrow c$, denoted by $corr(r)$, measures the relationship between the antecedent I and the consequent c. Formally, it is defined as $corr(r) = \frac{P(I \cup c)}{P(I)P(c)}$, where $P(I \cup c)$ is the occurrence frequency $supp(I \cup c)$ of $I \cup c$ across a set $\mathcal{D}$ of training cases. Analogously, $P(I)$ and $P(c)$ correspond to the occurrence frequencies of I and c in $\mathcal{D}$. If $corr(r) < 1$, r is negatively correlated. Instead, $corr(r) = 1$ denotes absence of correlation (i.e. I and c co-occur by chance), whereas $corr(r) > 1$ represents positive correlation. □

In highly imprecise learning settings, a class association rule r is interesting if it is positively correlated and also meets certain minimum requirements on its support and confidence. An associative classifier is a suitable disjunction of propositional `if-then` rules, that can be used for the classification of unlabeled cases.

Definition 2.6 (Associative classifier). An associative classifier $\mathcal{C}$ approximates the (unknown) discrete-valued case labeling function behind $\mathcal{D}$. The learnt approximation is represented as a disjunction $\mathcal{C} = \{r_1 \vee \ldots \vee r_k\}$ of interesting class association rules extracted from $\mathcal{D}$. □

An associative classifier $\mathcal{C}$ is used in section 3 to globally segment the whole training data, for the purpose of bringing to the surface those originally rare data, that becomes less rare within each resulting segment. Instead, in section 4, the CARs of $\mathcal{C}$ are viewed as properties local to each individual data case.

3. The *Global-to-Local* Supervised Learning Framework

We here discuss a *global-to-local* approach aimed to learn a hierarchical framework from the training cases $\mathcal{D}$, that consists of two classification levels. At the higher level, an associative classifier is built such that its component CARs meet some requirements on the minimum support and confidence. For each CAR $r \in \mathcal{C}$, the lowest level of the framework includes a local probabilistic generative model $P^{(r)}$ that allows to confirm or rectify r in the classification of an unlabeled case. The overall learning process is shown in figure 1. Given a database $\mathcal{D}$ of training cases (defined over a set $\mathcal{M}$ of items and a set $\mathcal{L}$ of class labels), the algorithm begins (at line 1) by extracting a set $\mathcal{R}$ of class association rules from $\mathcal{D}$ via the MINECARS search strategy.

HIERARCHICALLEARNING($\mathcal{M}$,$\mathcal{D}$,$\mathcal{L}$,τ)
Input: a finite set $\mathcal{M}$ of boolean attributes;
a training dataset $\mathcal{D}$;
a set $\mathcal{L}$ of class labels in $\mathcal{D}$;
and a support threshold τ;
Output: An associative classifier $\mathcal{C} = \{r_1 \vee \ldots \vee r_k\}$ and a set of local classifier $\mathcal{P}_{r_i}$;
1: $\mathcal{R} \leftarrow$ MINECARS$(\mathcal{M}, \mathcal{D}, \tau)$;
2: $\mathcal{R} \leftarrow$ ORDER$(\mathcal{R})$;
3: $\mathcal{C} \leftarrow$ PRUNE$(\mathcal{R})$;
4: **if** there are cases in $\mathcal{D}$ that are not covered by any rule within $\mathcal{C}$ **then**
5: $\mathcal{C} \leftarrow \mathcal{C} \cup \{r_d\}$;
6: **end if**
7: **for** each rule $r \in \mathcal{C}$, such that $r \neq r_d$ **do**
8: $\mathcal{P}^{(r)} \leftarrow$ TRAINLOCALCLASSIFIER(r);
9: **end for**
10: RETURN $\mathcal{C}$ and $\mathcal{P}^{(r)}$ for each $r \in \mathcal{C}$

Fig. 1. The hierarchical *global-to-local* learning framework

The rule set $\mathcal{R}$ is subsequently sorted (at line 2) according to the total order $\prec$, which is a refinement of the one in (Liu et al, 1998). Precisely, given two rules $r_i, r_j \in \mathcal{R}$, r_i precedes r_j, which is denoted by $r_i \prec r_j$, if (*i*) the confidence of r_i is greater than that of r_j, or (*ii*) their confidences are the same, but the support of r_i is greater than that of r_j, or (*iii*) both confidences and supports are the same, but r_i is shorter than r_j.

The learning process proceeds (at line 3) to distil a classifier $\mathcal{C}$ by pruning $\mathcal{R}$, which generally includes a very large number of CARs, that may overfit the training cases. For this purpose, the overfitting avoidance strategy presented in (Cesario et al, 2008) is exploited to reduce the complexity of the discovered CARs, while still improving their error rate. This is essentially accomplished via the removal of individual items and/or whole rules.

The resulting classifier $\mathcal{C}$ may leave some training cases uncovered. Therefore, a default rule $r_d : \emptyset \rightarrow c^*$ is appended to $\mathcal{C}$ (at line 5), such that its antecedent is empty and c^* is the majority class among the uncovered training cases.

As a remark, notice that, due to the total order $\prec$ enforced over $\mathcal{R}$, the associative classifier $\mathcal{C}$ is actually a decision list: each training case is classified by the first CAR in $\mathcal{C}$ that covers it. In other words, the CARs in $\mathcal{C}$ are mutually exclusive, i.e. a training case is covered by at most one rule of the classifier. As a consequence, the generic CAR $r : I \rightarrow c$ hereinafter covers the set of all those training cases that are not covered by any other CAR with higher precedence. More precisely, the definition of the coverage $\mathcal{D}_r$ of CAR r is refined into $\mathcal{D}_r = \{t \in \mathcal{D} | I \subseteq (t - \{class(t)\}) \wedge \nexists r' \in \mathcal{C} : r' \prec r, r' : I' \rightarrow c', I' \subseteq (t - \{class(t)\})\}$. Moreover, the addition to $\mathcal{C}$ (at line 5) of the default rule r_d ensures that $\mathcal{C}$ is also exhaustive, i.e. that every training case of $\mathcal{D}$ is covered by at least one CAR of $\mathcal{C}$.

Finally, for each CAR $r \in \mathcal{C}$ other than r_d, a local probabilistic model $\mathcal{P}^{(r)}$ is built (lines 7-9) over $\mathcal{D}_r$ to catch a better generalization of those globally rare cases/classes that become less rare within $\mathcal{D}_r$. This allows to refine the prediction from r with a local generative model that is better suited to deal with the local facets of rarity.

The MINECARS procedure is covered in subsection 3.1. The TRAINLOCALCLASSIFIER step is instead discussed in subsection 3.2, that also covers the classification of unlabeled cases (not reported in figure 1) in the context of two schemes for a tight integration between associative and local probabilistic classification.

3.1. Mining the Class Association Rules

MINECARS is an Apriori-based algorithm, adopted to mine positively-correlated CARs from the available training data $\mathcal{D}$. MINECARS combines into the basic Apriori algorithm (Agrawal and Srikant, 1994) two individually effective mechanisms, namely multiple minimum class support (Liu et al, 2000) and complement class support (Arunasalam and Chawla, 2006). Although both designed to deal with rarity in data, to the best of our knowledge, the joint effectiveness of such mechanisms has not yet been exploited. Figure 2 sketches the scheme of MINECARS algorithm, which divides into frequent itemset discovery (lines M1- M18) and CAR generation (lines M19- M26).

Frequent itemset discovery starts (at line M3) with C_2, a set of candidate 2-itemsets, including an item and a class label. At the generic iteration, MINECARS builds L_k, a set of frequent k-itemsets, from L_{k-1}. Two steps are performed to this purpose. The *join step* (at line M14) involves joining L_{k-1} with itself to yield C'_k, a collection of candidate k-itemsets. Notice that this requires joining pairs of frequent $k-1$-itemsets with identical class labels. The well-known *Apriori* property, according to which an unfrequent itemset cannot have frequent supersets, is then used (at line M15) to drop from C'_k those k-itemsets with at least one $k-1$-subset that is not in L_{k-1}. The *support counting step* (lines M5- M12) involves counting the occurrences of the surveyed candidate itemsets in C_k by scanning the training data $\mathcal{D}$. Those candidates whose support exceeds a class-specific threshold are considered to be frequent and retained within L_k. The level-wise search halts when no more frequent itemsets can be discovered.

Multiple minimum class support (Liu et al, 2000) is employed at line M13 to the purpose of automatically adjusting the global minimum support threshold τ provided by the user to minimum support threshold specific for each class. Essentially, the generic candidate itemset c is frequent if its support is over $\tau \cdot \sigma(class(c))$, the minimum support threshold for $class(c)$. Multiple minimum class support implements a first stage of focused pruning, that dynamically assigns a higher minimum support threshold to majority classes (which prevents from yielding several overfitting rules) and a lower minimum support threshold to minority classes (which enforces the generation of an appropriate number of rules).

Complement class support (Arunasalam and Chawla, 2006) is instead used in the CAR generation stage, to avoid the specification of a global minimum confidence threshold. In particular, a specific property of complement class support (shown in (Arunasalam and Chawla, 2006)) is exploited at line M22 to automatically identify a class-specific minimum confidence threshold. According to such a property, a rule $r : I \rightarrow c$ is such that I and c are positively correlated if and only if $conf(I \rightarrow c) > \frac{\sigma(c)}{|\mathcal{D}|}$, where $\sigma(c)$ is the overall number of occurrences of class c in $\mathcal{D}$. Therefore, the CARs whose confidence exceeds (at line M22) the minimum threshold corresponding to their targeted class are guaranteed to be positively correlated. Thus, both confidence and positive correlation between rule components can be verified without additional parameters or further correlation analysis.

The dynamic selection of a class-specific minimum confidence threshold acts essentially a second stage of focused pruning, that ensures the discovery of accurate rules targeting the rare classes and still avoids the generation of an overwhelming number of rules from the predominant classes.

MINECARS($\mathcal{M}$,$\mathcal{D}$,τ)
Input: a finite set of boolean attributes $\mathcal{M}$;
a training dataset $\mathcal{D}$;
and a support threshold τ;
Output: a set $\mathcal{R}$ of class association rules;
/* Frequent itemset discovery */
M1: $I \leftarrow \emptyset$, $k \leftarrow 2$;
M2: Let $\mathcal{L}$ be the set of class labels in $\mathcal{D}$;
M3: Let $C_2 \leftarrow \{c | c = \{l, i\}$ where $l \in \mathcal{L}, i \in \mathcal{M}\}$;
M4: **while** $C_k \neq \emptyset$ **do**
M5: **for** each candidate itemset $c \in C_k$ **do**
M6: $supp(c) \leftarrow 0$;
M7: **end for**
M8: **for** $t \in \mathcal{D}$ **do**
M9: **for** $c \in C_k$ such that $c \subseteq t$ **do**
M10: $supp(c) \leftarrow supp(c) + \frac{1}{|\mathcal{D}|}$;
M11: **end for**
M12: **end for**
M13: $L_k \leftarrow \{c \in C_k | supp(c) > \tau \cdot supp(class(c))\}$;
M14: $C'_{k+1} \leftarrow \{c_i \cup c_j | c_i, c_j \in L_k \wedge class(c_i) = class(c_j) \wedge |c_i \cup c_j| = k+1\}$;
M15: $C_{k+1} \leftarrow \{c \in C'_{k+1} | \forall c' \subset c$ such that $|c'| = k$ it holds that $c' \in L_k\}$;
M16: $k \leftarrow k + 1$;
M17: **end while**
M18: $\mathcal{I} \leftarrow \cup_k L_k$;
/* CAR generation */
M19: $\mathcal{R} \leftarrow \emptyset$;
M20: **for** each frequent itemset $I \in \mathcal{I}$ **do**
M21: create rule $r : I - class(I) \rightarrow class(I)$;
M22: **if** $conf(r) > \frac{\sigma(class(I))}{|\mathcal{D}|}$ **then**
M23: $\mathcal{R} \leftarrow \mathcal{R} \cup \{r\}$;
M24: **end if**
M25: **end for**
M26: RETURN $\mathcal{R}$;

Fig. 2. The process for mining class association rules from data with rarity

3.2. Training Local Classifiers

The individual CARs of the associative classifier $\mathcal{C}$ predict classes for unlabeled cases based on global statistics, that take into account the regularities across the whole training data $\mathcal{D}$. This makes targeting rare classes problematic. Therein, it is reasonable to assume that the class assigned to the unlabeled cases covered by the generic CAR r should be influenced more strongly by the classes of the training cases local to r (i.e., falling within $\mathcal{D}_r$) and less strongly by the classes of farther training cases (covered by CARs other than r). According to this intuition, the prediction of each CAR $r \in \mathcal{C}$ can be refined by associating the latter with a local probabilistic generative model $\mathcal{P}^{(r)}$, trained over the regularities across the training cases falling within $\mathcal{D}_r$. In principle, such regularities are likely to be more descriptive of those globally rare classes that become less rare within $\mathcal{D}_r$. Consequently, the individual $\mathcal{P}^{(r)}$ can be involved into the classification process for more accurately dealing with the corresponding forms of rarity. As a matter of fact, coupling $\mathcal{C}$ with local probabilistic classification models is useful to improve the classification performance of $\mathcal{C}$ both in the surroundings of the decision boundaries of its CARs as well as within the inner areas of CARs' decision regions (wherein classes

other than the ones associated to the whole regions are likely to meaningfully influence the classification of nearby unlabeled cases).

In the following, we adopt two different probabilistic generative models based, respectively, on the naïve Bayes and nearest neighbor classification models. Precisely, naïve Bayes naturally allows to incorporate the effects of locality on classes and cases in terms of, respectively, class priors and item posteriors. To elucidate, an unlabeled case $I \subset \mathcal{M}$ is assigned by the generic generative model $\mathcal{P}^{(r)}$ to the class $c \in \mathcal{L}$ with highest posterior probability

$$\mathcal{P}^{(r)}(c|I) \triangleq p(c|I,r) = \frac{\prod_{i \in I} p(i|c,r)p(c|r)}{\sum_{\overline{c} \in \mathcal{L}} \prod_{i \in I} p(i|\overline{c},r)p(\overline{c}|r)}$$

Locality influences factors $p(c|r)$'s and $p(i|c,r)$'s, whose values are estimated by computing $p(c)$ and $p(i|c)$ over $\mathcal{D}_r$, and allows to better value rare cases/classes. Indeed, if a significant extent of some form of rarity falls within $\mathcal{D}_r$, the corresponding cases/classes are obviously less rare than in $\mathcal{D}$ and, hence, factors $p(c)$'s and $p(i|c)$'s are accordingly higher (w.r.t. their values in $\mathcal{D}$). Dually, $p(c)$'s and $p(i|c)$'s are sensibly lower, if the density of that form of rarity within $\mathcal{D}_r$ is much lower than in $\mathcal{D}$. However, this is acceptable, since most of that form of rarity is still captured within some other region(s). An inconvenient behind the adoption of naïve Bayes as the underlying model for local probabilistic classifiers is their performance degrade (e.g. accuracy loss) due to the violation of the attribute independence assumption. To alleviate such an issue, the weaker attribute independence assumption postulated in AODE (Webb et al, 2005) can be plugged into the above formulation, that simply refines naïve Bayes by considering each attribute dependent upon at most n other attributes in addition to the class. This is more realistic in practical applications and is empirically shown in section 5 to yield a better performance. We omit the formal discussion on the n-dependence estimation technique behind AODE and refer to (Webb et al, 2005).

The nearest neighbor model can be alternatively used as a local model to compute probabilities $\mathcal{P}^{(r)}(c|I)$ from the distribution of classes within $\mathcal{D}_r$ through the generative approach below

$$\mathcal{P}^{(r)}(c|I) \triangleq \frac{\sum_{I' \in \mathcal{D}_r} w_{I'} p(c|I')}{\sum_{\overline{c} \in \mathcal{L}} \sum_{I' \in \mathcal{D}_r} w_{I'} p(\overline{c}|I')}$$

The above is essentially a probabilistic re-formulation of a distance-weighted voting scheme, in which each neighbor I' votes for the class that should be assigned to I. The vote from the generic neighbor I' is suitably weighted by a corresponding factor $w_{I'}^{(r)}$, which takes into account the actual distance between I' and I. Formally,

$$w_{I'} = \frac{e^{-d^2(I,I')}}{\sum_{I' \in \mathcal{D}_r} e^{-d^2(I,I')}}$$

where $d(I,I')$ is any suitable function that defines a notion of distance between I and I'.

Notice that, whatever the distance between cases, the chosen weight-definition attributes higher influences to those neighbors in $\mathcal{D}_r$ that are actually closest to I.

Two alternative approaches for refining the predictions from the associative classifier $\mathcal{C}$ through the local probabilistic generative models $\mathcal{P}^{(r)}$'s are discussed next.

Local priors and local instance posteriors. The idea is to reformulate a generative approach to classification which spans into local generative models. Given an unlabeled case I, we introduce a space of events $\Omega = \{e_r | r \in \mathcal{C}\}$ related to the classification of I via an associative classifier $\mathcal{C}$. More precisely, the individual event e_r corresponds to the coverage of I through a corresponding CAR $r \in \mathcal{C}$. The exclusiveness and exhaustiveness of the CARs in $\mathcal{C}$ imply, respectively, the mutual exclusiveness and the collective exhaustiveness of the events in Ω. Therefore, it is possible to employ the well-known law of total probability to define a joint probability distribution over unlabeled cases and class labels as shown below

$$p(c, I) = \sum_{e_r \in \Omega} p(c, I, r) = \sum_{e_r \in \Omega} p(c, I|r)p(r) = \sum_{e_r \in \Omega} \mathcal{P}^{(r)}(c|I)p(I|r)p(r)$$

The interpretation of the terms within the above formula is provided next. $p(I|r)$ represents the compatibility of I with the rule r. We choose to model $p(I|r)$ as the relative number of items that I shares with r: intuitively, the number of (mis)matches represents the closeness of I to the region bounded by r. $\mathcal{P}^{(r)}(c|I)$ denotes the probability associated with c by the local naïve Bayes classifier $\mathcal{P}^{(r)}$ trained over $\mathcal{D}_r$. $p(r)$ indicates the support $supp(r)$ of CAR r and weights its contributions to $p(c, I)$ by the relative degree of rarity of its antecedent and consequent.

Finally, the probability of class c given the unlabeled case I can be formalized as the following generative model

$$p(c|I) = \frac{p(c, I)}{\sum_{\overline{c} \in \mathcal{L}} p(\overline{c}, I)}$$

Cumulative rule effect. A stronger type of interaction between global and local effects can be injected into the classification process, if the predictions from a CAR r and unrelated local generative model $\mathcal{P}^{(r')}$ (with $r \neq r'$) are compared for selecting the most confident one. The overall approach is sketched in figure 3. Precisely, the generic unlabeled case $I \subset \mathcal{M}$ is presented to the associative classifier $\mathcal{C}$ and the first CAR $r : I \rightarrow c$ (in the precedence order $\prec$ enforced over $\mathcal{C}$) is chosen (at line N1). If r does not cover I, it is skipped and the next rule is recursively taken into account (at line N20). Otherwise, r is used for prediction. However, its target class c is not directly assigned to I. Rather, the local probabilistic generative model $\mathcal{P}^{(r)}$ corresponding to r is exploited to produce a possibly more accurate prediction (at line N4). Some tests are performed to identify the more confident prediction (lines N9- N15). If both counterparts agree or one is deemed to be more reliable than the other one, the better prediction (in terms of class-membership probability distribution) is returned (lines N10 and N12). Otherwise, in the absence of strong evidence to reject the prediction from $\mathcal{P}^{(r)}$ (which is in principle preferable to r, being more representative of the local regularities that may come from globally rare cases/classes that fall within $\mathcal{D}_r$), r is skipped in favor of the next CAR $r' \in \mathcal{C}$ covering I (at line N14). To this point, if $\mathcal{P}^{(r')}$ predicts I more confidently than $\mathcal{P}^{(r)}$ (at line N5), the probability distribution from $\mathcal{P}^{(r')}$ replaces the current best distribution yielded by $\mathcal{P}^{(r)}$ (at line N6) and the choice of a better prediction is hence made between r' and $\mathcal{P}^{(r')}$. In the opposite case, the choice involves r' and the current best distribution $\mathcal{P}^{(r)}$. If no prediction is clearly eligible as the most confident throughout the search, the process halts when the default rule is met and the current best distribution is returned (at line N17). Notice that the sofar best class-membership

PREDICTION($\mathcal{C}$,I,$p_1, \ldots, p_k$)
Input: An associative classifier $\mathcal{C}$;
an unlabeled case $I \subset \mathcal{M}$;
Output: the class distribution for I;
N1: select the first rule $r : I' \rightarrow c_h$ in sequence within $\mathcal{C}$;
N2: **if** r covers I (i.e. $I' \subseteq I$) **then**
N3: **if** $|\mathcal{C}| > 1$ (i.e. r is not the default rule) **then**
N4: let $\overline{p}_i = \mathcal{P}^{(r)}(c_i|I) \cdot acc^{(c_i)}\left(\mathcal{P}^{(r)}\right), \forall i = 1, \ldots, k$;
N5: **if** $max_i(\overline{p}_i) > max_i(p_i)$ **then**
N6: let $p_i = \overline{p}_i, \forall i = 1, \ldots, k$;
N7: **end if**
N8: let $p^* = max_i(p_i)$ and $i^* = argmax_i(p_i)$ and $p = \sum_i p_i$;
N9: **if** $acc^{(c_h)}(r) < p^*$ **then**
N10: RETURN the distribution $(p_1/p, \ldots, p_k/p)$;
N11: **else if** $i^* = h$ **or** $acc^{(c_h)}(r) > \frac{p*}{p}$ **then**
N12: RETURN the distribution $(acc^{(c_1)}(r), \ldots, acc^{(c_k)}(r))$;
N13: **else**
N14: PREDICTION($\mathcal{C} - \{r\}$,I,$p_1, \ldots, p_k$);
N15: **end if**
N16: **else**
N17: RETURN the distribution $(p_1/p, \ldots, p_k/p)$;
N18: **end if**
N19: **else**
N20: PREDICTION($\mathcal{C} - \{r\}$,I,$p_1, \ldots, p_k$);
N21: **end if**

Fig. 3. The scheme of the cumulative rule effect

probability distribution is remembered throughout the consecutive stages of the search process via the input arguments $p_1, \ldots, p_k$ (such arguments are individually set to 0 at the beginning of the search process). A key aspect of the overall search process is represented by the criteria adopted to choose the more confident prediction between the ones from a CAR r_h and a local probabilistic generative model $\mathcal{P}^{(r_i)}$. Accuracy is used as a discriminant between the alternatives. In particular, the accuracy $acc^{(c)}\left(\mathcal{P}^{(r_i)}\right)$ is the percent of cases in $\mathcal{D}^{(r)}$ correctly predicted by $\mathcal{P}^{(r_i)}$ as belonging to class c. The accuracy $acc^{(c)}(r_h)$ of a CAR r_h predicting class c is its confidence $conf\,(r_h)$. When comparing the accuracies of a CAR r_h and a local probabilistic generative model $\mathcal{P}^{(r_i)}$ there are four possible outcomes.

1. $\mathcal{P}^{(r_i)}$ is clearly deemed more reliable than r_h (at line N9), if the weighted accuracy of the former, p^*, is greater than the accuracy of the latter.
2. r_h is preferred to $\mathcal{P}^{(r_i)}$ (at line N11) if the accuracy of the former is greater than or equal to the weighted accuracy of the latter and both agree anyhow.
3. r_h is preferred to $\mathcal{P}^{(r_i)}$ (again at line N11) if its accuracy is much greater than the weighted accuracy of $\mathcal{P}^{(r_i)}$. Therein, $\frac{p^*}{p} > p*$ is a prudential threshold, that represents the normalized weighted accuracy from $\mathcal{P}^{(r_i)}$. In practice, r_h is actually preferable to $\mathcal{P}^{(r_i)}$ if its accuracy exceeds $\frac{p^*}{p}$.
4. There is no strong evidence (at line N16) to reject either r_h or $\mathcal{P}^{(r_i)}$ when the accuracy of r_h lies in the interval $(p*, \frac{p^*}{p})$. In such a case, r is skipped and the search proceeds to considering the next CAR in the associative classifier $\mathcal{C}$ that covers I (through the recursive call at line N14).

4. The *Local-to-Global* Supervised Learning Framework

We here propose a *local-to-global* learning framework, that uses suitable features local to a data case, for predicting the global conditional probability of classes given the case. Features are a sort of declarative mechanism for specifying aspects of data cases that are relevant, to some extent, towards classification into the individual classes. The relevance of a feature with respect to a certain class determines the weight of that feature on the discrimination of the particular class. Thus, the recognition of minority classes can be addressed by identifying specific features that are highly representative of such classes.

The starting point is the observation that the training data $\mathcal{D}$ generally provides partial information on the associations between data cases and corresponding class labels. The latter is especially true in imprecise domains because of rarity. This suggests that the conditional probability distribution of classes should minimize commitment, i.e. fit the evidence observable in $\mathcal{D}$ and still be as uniform as possible in the prediction of whatever is not observable in $\mathcal{D}$. Such a conditional probability distribution represents the most unbiased assignment of class probabilities complying with the observable evidence. Any other probabilistic assignment would be biased, i.e. would assume the availability of arbitrary information that is not present in $\mathcal{D}$.

To elucidate, consider an hypothetical four-class classification setting, where $\mathcal{L} = \{c_1, c_2, c_3, c_4\}$. Classes c_1 and c_2 are rare, whereas c_3 and c_4 are predominant. Assume that an exploratory analysis of $\mathcal{D}$ reveals that a certain itemset $I \subset \mathcal{M}$ appears within classes c_1 and c_2 with a frequency that amounts to, respectively, 50% and 30% of the overall number of its occurrences. I can be viewed as a data feature and the statistical observations concerning I can be stated as constraints for the conditional probability distribution in order for the latter to agree with the empirical evidence. When a newly arrived case I' is presented to the conditional probability distribution for classification, there are two possibilities. If I' includes I (i.e. $I \subseteq I'$), the conditional probability distribution provides the assignments $p(c_1|I') = 0.5$ and $p(c_2|I') = 0.3$. The remaining 0.2 of the probability mass is uniformly distributed between classes c_3 and c_4 (in the absence of any further specific information on this aspect), so that $p(c_3|I') = 0.1$ and $p(c_4|I') = 0.1$. Notably, I is inherently characteristic of the rare classes c_1 and c_2 and its adoption as a data feature allows for a proper discrimination of such classes. If instead I' does not contain I, the conditional probability distribution assumes (in the absence of any further evidence observable in $\mathcal{D}$) maximal ignorance and, hence, predicts each of the four possible classes with uniform probability, i.e. $p(c_1|I') = p(c_2|I') = p(c_3|I') = p(c_4|I') = 0.25$. In this manner, the conditional probability distribution agrees with the observed evidence in $\mathcal{D}$ and still avoids assumptions on whatever is unknown.

The *local-to-global* approach relies on statistical modeling and learning to fit the evidence in $\mathcal{D}$. For this purpose, the training data is used to identify a set of features useful for classification. The individual features are then employed to specify as many constraints for the conditional probability distribution to learn. The generic constraint essentially forces the expected value that the conditional probability distribution assigns to some corresponding feature to be the same as the expected value of that feature empirically observed in $\mathcal{D}$. In general, the space of features can be potentially very large. In these cases, computing the optimal conditional probability distribution as a closed form solution that meets all the specified constraints is prohibitive. Maximum entropy model (Berger et al, 1996) provides an expressive and powerful mathematical framework for iteratively computing the required distribution. It is also used to elegantly and seamlessly integrate two established methods from the fields of machine learning and data mining. On the machine learning side, discriminative learning is used to directly compute the conditional probability distribution of the classes, given an unseen case.

The main difference with respect to generative learning, that would instead model a joint probability distribution over classes and cases, is that discriminative learning allows to better fit the training data by carefully setting the distribution parameters. On the data mining side, associative classification provides the space of features to which the training and newly arrived data is mapped.

4.1. Modeling Data Evidence through CARs, Features and Constraints

In the proposed *local-to-global* learning framework, features are associated to the individual CARs of an associative classifier formed as described in subsection 3.1 (no further post-pruning is applied to these CARs). Therein, let $\mathcal{C}$ be an associative classifier. The space of features $\mathcal{F} = \{f_{r_i} | r_i \in \mathcal{C}\}$ is essentially a finite set of real-valued indicator functions f_{r_i}, each of which is associated to a corresponding CAR $r_i \in \mathcal{C}$. Assume that the generic r_i has the implicative structure $r_i : I' \rightarrow c'$, with $I' \subset \mathcal{M}$ and $c' \in \mathcal{L}$. Moreover, let $I \subset \mathcal{M}$ denote a data case and $c \in \mathcal{L}$ represent any class label. The generic feature f_{r_i} is defined as the following indicator function

$$f_{r_i:I' \rightarrow c'}(I, c) = \begin{cases} 1 & \text{if } I' \subseteq I \wedge c' = c \\ 0 & \text{otherwise} \end{cases}$$

The individual feature $f_{r_i:I' \rightarrow c'}$ is said to be *local* to I if $f_{r_i:I' \rightarrow c'}(I, c) > 0$. Training and newly arrived cases can hence be represented as suitable configurations of local features, which are useful for classification. Intuitively, the interpretation of CAR r_i is that I' can be viewed as a sort of contrast set (Bay and Pazzani, 2001) for the targeted class c', i.e. as a co-occurrence of items that is inherently characteristic of c', since its distribution in $\mathcal{D}$ is meaningfully associated with the class c'. Therefore, if r_i covers I (i.e. $I' \subseteq I$) and the class c considered for I coincides with the class c' targeted by r_i, c' is an eligible class for I. Therein, a measure of the suitability of c' for I is provided by the value of $f_{r_i:I' \rightarrow c'}$.

Actually, features are normalized so that their sum amounts to 1. Here, we specify that the generic feature $f_{r_i:I' \rightarrow c'}(I, c)$ is normalized into the corresponding

$$f'_{r_i:I' \rightarrow c'}(I, c) = \frac{f_{r_i:I' \rightarrow c'}(I, c)}{\sum_{r_i} f_{r_i:I' \rightarrow c'}(I, c)}$$

and additionally highlight that, for simplicity, the original notation $f_{r_i:I' \rightarrow c'}(I, c)$ is still maintained in the ongoing discussion to mean $f'_{r_i:I' \rightarrow c'}(I, c)$.

Features are the basic building block for specifying constrains. These are necessary to make the required conditional probability distribution fit the observed evidence in $\mathcal{D}$. To elaborate, the empirical evidence relative to each feature f_{r_i} is summarized into $E_{\mathcal{D}}(f_{r_i})$, which is the expected value of f_{r_i} observed in $\mathcal{D}$. Its definition is

$$E_{\mathcal{D}}(f_{r_i}) = \sum_{t \in \mathcal{D}} p_{\mathcal{D}}(t) f_{r_i}(t - class(t), class(t))$$

where $p_{\mathcal{D}}$ is the observed occurrence frequency of t in $\mathcal{D}$, i.e. $p_{\mathcal{D}}(t) = \frac{1}{|\mathcal{D}|}$. Constraints force the required conditional probability distribution $\mathcal{P}$ to agree with the feature expectations observed in $\mathcal{D}$. In other words, for each feature f_{r_i}, a corresponding constraint is specified that equates the expected value that $\mathcal{P}$ assigns to f_{r_i} to the expected value $E_{\mathcal{D}}(f_{r_i})$ observed in $\mathcal{D}$. With respect to the generic feature f_{r_i}, the expected value of f_{r_i} due to $\mathcal{P}$ can be approximated as shown below:

$$
\begin{aligned}
E(f_{r_i}) &= \sum_{I \subset \mathcal{M}, c \in \mathcal{L}} p(I,c) f_{r_i}(I,c) \\
&= \sum_{I \subset \mathcal{M}} p(I) \sum_{c \in \mathcal{L}} \mathcal{P}(c|I) f_{r_i}(I,c) \\
&\approx \sum_{I \subset \mathcal{M}} p_{\mathcal{D}}(I) \sum_{c \in \mathcal{L}} \mathcal{P}(c|I) f_{r_i}(I,c)
\end{aligned}
$$

where the (unknown) prior probability distribution of cases $p(\cdot)$ is approximated by the empirical distribution $p_{\mathcal{D}}(\cdot)$. The term $p_{\mathcal{D}}(I)$ approximates $p(I)$ by the occurrence frequency of I in $\mathcal{D}$.

Finally, restricting $\mathcal{P}$ to have the same feature expectations as the ones observed in $\mathcal{D}$ requires setting the following constraints

$$E(f_{r_i}) = E_{\mathcal{D}}(f_{r_i}) \text{ for each } f_{r_i} \in \mathcal{F}$$

The above restrictions exclude from further consideration all those conditional probability distributions, that do not accord with the observed feature expectations.

In principle, there are infinitely many conditional probability distributions consistent with the specified constraints. The maximum entropy principle suggests to choose the conditional probability distribution $\mathcal{P}$ that fits the constraints (i.e. agrees with the observed evidence) and maximizes entropy for those cases that are not subject to the constraints. These latter cases are hence predicted to be members of the distinct classes with the most uniform probability distribution.

The mathematical derivation of the required conditional distribution $\mathcal{P}$ as well as the estimation of its parameters are beyond the scope of this manuscript. The interested reader is referred to (Berger et al, 1996) for an exhaustive coverage.

5. Evaluation

We conduct a systematic experimental study devoted to understanding whether the proposed hierarchical classification scheme exhibits improvement in classification performance with respect to established competitors. To this purpose, the comparative evaluation is carried out over some standard datasets. In particular, we use some datasets chosen from the UCI KDD repository (Asuncion and Newman, 2007), with high class imbalance. Also, we test our approach over the KDD99 intrusion detection dataset, named `kdd99`. The latter is an extremely unbalanced dataset, wherein low-frequency classes are characterized by noise. A further non-publicly available test dataset, `fraud`, is a real-life fraud detection dataset, with a very low class separability.

Notice that, even if the number of tuples belonging to the class *u2r* is very limited, `kdd99` is a standard benchmark dataset utilized for classification with unbalanced classes, and the challenge there is in particular to detect the minority classes (including the class *u2r*). However the improvements in our methodology is testified also by the `fraud` dataset, where the differences in frequency are not so skewed.

Experiments consists in comparisons against several established rule-based and associative classifiers. The selected rule-based competitors are Ripper (Cohen, 1995) and PART (Frank and Witten, 1998), while the associative ones include CBA (Liu et al, 1998) and CMAR (Xin and Han, 2003). In particular, we exploited the implementa-

tions of CBA and CMAR in (Coenen, 2004). All tests are conducted on an Intel Itanium processor with 4Gb of memory and 2Ghz of clock speed.

All numeric attributes in the selected datasets are suitably discretized prior to the application of the devised schemes. The adopted discretization strategy partitions the values of each numeric attribute into natural clusters, via model-based clustering. The idea is to view the values of a numeric attribute as the result of a statistical generative process, which is modeled through a mixture of univariate Gaussian distributions. For each numeric attribute, the choice of the most appropriate number of clusters (i.e. distinct Gaussian distributions in the mixture model) is performed by letting such number range from 1 up to a certain maximum, which is fixed to 16 in our experimental setting. The discretization of each numeric attribute into any number of clusters in the foresaid range is then assessed through 5-fold cross validation. More precisely, 4 folds of attribute values are used to estimate the values of the parameters in the hypothesized mixture model (i.e. the mean and standard deviation for each Gaussian distribution as well as the weights of the individual distributions) by means of the well-known EM algorithm (McLachlan and Peel, 2000). The remaining fold is employed to evaluate discretization quality. This latter step involves employing the estimates of model parameters to compute the likelihood of the attribute values in the test fold. Eventually, the number of clusters chosen to partition the values of the generic numeric attribute is the one with maximum average likelihood on the test fold and, thus, the values of the attribute are replaced by the label of the cluster to which they belong with highest probability.

The execution of the selected classifiers is reiterated several times, under different parameter configurations and the result of the individual execution were averaged through leave-one-out method. For each classifier, we next report the results corresponding to the best parameter configuration.

Our schemes simply require the specification of a global minimum support. Due to the adoption of minimum class support (Liu et al, 2000), such threshold is automatically adjusted to become a class specific threshold. In particular, we fixed the global support threshold to 20%, which is transparently adjusted to be, within the individual class in the data at hand, the 20% of the frequency of that class. The exploitation of complement class support (Arunasalam and Chawla, 2006) permits to avoid specifying a minimum confidence threshold.

We compare the approaches using accuracy, some meaningful ROC curves and the Area Under the Curve (AUC) relative to the minority class. Tables 1 and 2 display the results. Within the tables, competitors are numbered from (1) to (4). Precisely, (1) indicates Ripper, (2) corresponds to PART, while (3) and (4) stand for CBA and CMAR, respectively. Our schemes are instead numbered from (5) to (11). More specifically, (5) and (6) indicate naive Bayesian smoothing (respectively through local priors or cumulative effect). (7) and (8) stand for nearest-neighbor smoothing (respectively, through local priors or cumulative effect). (9) and (10) are AODE smoothing (respectively, through local priors or cumulative effect). Finally, (11) represents the maximum entropy approach.

The results clearly state that the combination of associative classification and probabilistic smoothing is at least as accurate as the seminal rule-based classifiers chosen for the comparison. In many cases, however, (5) and (11) achieve improvements in accuracy, reported in bold within table 1, that are statistically significant according to the t-test. In addition, a deeper analysis reveals that the response versus the classes of interest is strongly improved. Such an improvement can be appreciated by looking at the details of the individual datasets. To elucidate, we report in table 5 the confusion matrices originated by (1) and (9) over the `german-credit` dataset. Notice that

Table 1. Classification accuracy (expressed in %)

Dataset	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**
anneal	98.26	98.25	92.81	96.33	98.53	98.53	**99.41**	99.10	98.43	98.43	66.70
balance-scale	80.30	**83.17**	68.81	68.49	81.40	81.04	81.98	82.27	80.27	80.30	63.52
breast-cancer	71.45	69.41	69.20	67.67	70.34	70.34	72.25	68.62	72.30	72.30	**76.02**
horse-colic	85.10	84.37	81.62	83.96	82.56	82.56	82.74	82.74	83.20	83.20	**85.21**
credit-rating	85.16	84.45	81.74	83.76	80.48	80.48	81.36	81.36	**85.90**	**85.90**	85.32
german_credit	72.21	70.54	73.10	73.34	74.03	74.03	71.00	71.00	**74.87**	**74.87**	69.67
pima_diabetes	75.18	73.45	**77.87**	73.03	73.31	73.31	70.48	70.48	75.02	75.02	64.39
Glass	66.78	68.75	72.69	**74.23**	58.94	59.17	68.15	68.15	67.48	67.15	67.32
cleveland-14-heart-diseas	79.95	78.00	82.12	75.12	81.29	81.29	77.49	77.49	81.15	81.01	**90.54**
hungarian-14-heart-diseas	79.57	81.14	82.06	79.69	81.24	81.24	77.64	77.64	82.62	82.38	**86.70**
heart-statlog	78.70	77.33	82.59	84.19	80.41	80.41	75.74	75.74	78.96	78.96	**88.70**
hepatitis	78.13	79.80	79.89	81.08	**81.22**	**81.22**	80.34	80.34	81.10	81.10	80.63
ionosphere	89.16	**90.83**	87.89	89.74	82.85	82.85	89.47	89.47	88.30	88.30	75.21
labor	83.70	77.73	86.67	88.77	84.60	84.60	85.50	85.50	87.13	87.13	**100.00**
lymphography	76.31	76.37	81.18	**89.59**	78.38	78.38	78.31	77.92	78.00	78.08	88.54
sick	98.29	**98.62**	97.51	97.64	98.25	98.25	98.58	98.58	98.39	98.39	97.64
sonar	73.40	77.40	80.00	82.78	75.28	75.28	78.39	78.39	73.79	73.79	**96.63**
fraud	93.07	93.02	80.82	90.52	91.78	91.79	93.05	92.96	92.61	92.61	**93.27**
kdd99	96.61	**96.98**	94.65	94.63	95.98	95.98	96.78	96.73	96.65	96.65	92.85

Table 2. Area Under the Curve

Dataset	**(1)**	**(2)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**
anneal	0.76	0.88	**0.93**	**0.93**	0.90	0.74	**0.93**	**0.93**	0.90
balance-scale	0.86	0.92	0.94	0.94	**0.96**	0.95	0.87	0.87	0.67
breast-cancer	0.60	0.59	0.67	0.67	0.63	0.62	0.69	0.69	**0.75**
horse-colic	0.83	0.86	0.85	0.85	0.81	0.81	0.88	0.88	**0.95**
credit-rating	0.87	0.88	0.88	0.88	0.82	0.82	0.93	0.93	**0.94**
german_credit	0.63	0.67	0.77	0.77	0.66	0.66	**0.78**	**0.78**	0.71
pima_diabetes	0.72	0.78	0.78	0.78	0.70	0.70	**0.79**	**0.79**	0.76
Glass	0.80	0.79	0.80	0.80	0.77	0.77	**0.81**	0.80	0.76
cleveland-14-heart-diseas	0.81	0.80	0.88	0.88	0.78	0.78	0.90	0.89	**0.97**
hungarian-14-heart-diseas	0.78	0.86	0.88	0.88	0.78	0.78	0.90	0.90	**0.92**
heart-statlog	0.80	0.78	0.86	0.86	0.77	0.77	0.81	0.81	**0.96**
hepatitis	0.62	0.78	0.80	0.80	0.68	0.68	0.84	0.84	**0.99**
ionosphere	0.89	0.89	0.90	0.90	0.91	0.91	0.90	0.90	**0.97**
labor	0.82	0.73	0.86	0.86	0.83	0.83	0.95	0.95	1.00
lymphography	0.40	0.64	0.56	0.56	0.66	0.68	**0.98**	0.89	0.94
sick	0.94	0.95	**0.97**	**0.97**	0.95	0.95	0.96	0.96	0.90
sonar	0.75	0.79	0.80	0.80	0.80	0.80	0.77	0.77	1.00
fraud	0.97	0.97	0.90	0.90	0.97	0.97	**0.98**	0.97	0.97
kdd99	0.98	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	0.94

the probabilistic smoothing recovers 39 tuples to the minority class, thus allowing to achieve higher precision.

A further analysis of the results obtained over the `fraud` and the `kdd99` datasets provides an in-depth into the effects of smoothing. Figure 4 shows the ROC curves relative to (1), (2), (5), (7) and (9). There is an evident improvement in the underlying area with respect to the competitors (1) and (2), whose trends are plotted in red. Results with the `kdd99` dataset are even more surprising, and in particular with the `u2r` class, as shown in figure 4, that represents the curves relative to the schemes (1), (2) and (9).

Predicted ->	good	bad
good	607	93
bad	155	145

AODE local priors (9)

Predicted ->	good	bad
good	611	89
bad	194	106

Ripper (1)

Table 3. A comparison between the confusion matrices yielded by AODE local priors (9) and Ripper (1)

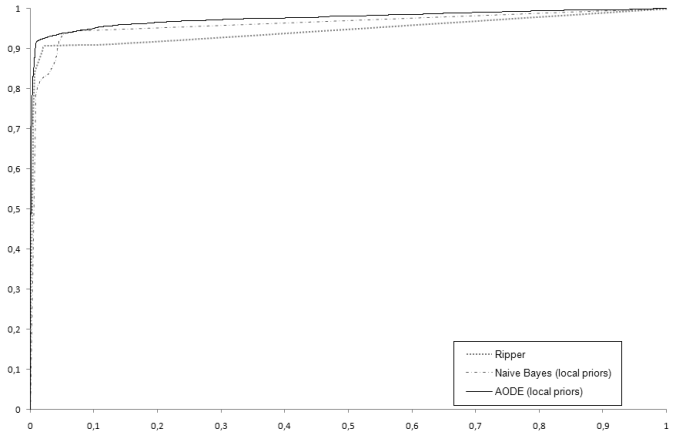

Fig. 4. ROC curve for the minority class of the `fraud` data set

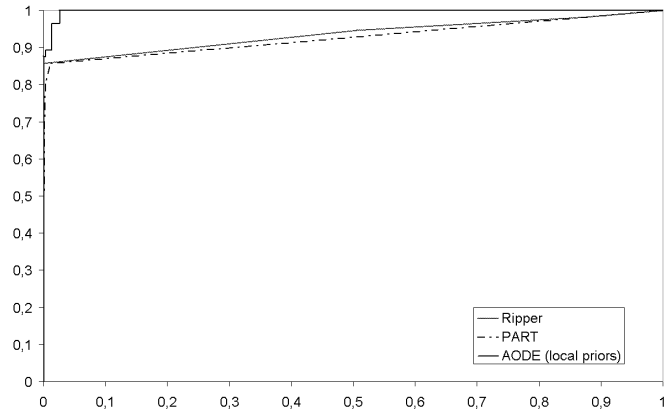

Fig. 5. ROC curve for the minority `u2r` class within the `kdd99` dataset

The `u2r` class is made of 56 tuples (out of 500K), and still the probabilistic adjustment is capable of recovering some problematic cases.

The ability of the approaches at dealing with the classes is compared in table 2, which tabulates the average of the AUC values over the classes within the selected datasets. Overall, the devised schemes outperform the competitors by exhibiting a significantly improved performance (i.e. a considerable increase in the area under the ROC curve) across all classes within the distinct datasets and, in particular, with `hepatitis`, `lymphography` and `fraud`, where the improvement is over 10%. As witnessed by the graphs in figures 4 and 5, such an overall improvement is primarily obtained on the minority classes.

A detailed comparison of the above described approaches is provided in tables 5, 6, 7 and 8 in the appendix. In these tables, for each class label, we report precision, recall, f-measure and AUC to better highlight the effects of the proposed techniques on highly unbalanced datasets.

Finally, the table 4 shows shows the running times of the representative schemes in

Table 4. Running times (in seconds) vs. data size

Dataset	Size	# Attributes	Ripper (1)	MRNB (5)	MaxEnt (11)
german_credit	1000	21	0.13	0.17	14
sick	3772	28	0.33	0.88	29
fraud	45442	76	321	436	795
kdd99	494020	42	68	152	396

the above mentioned tables, for datasets of increasing size. There is an obvious overhead in the proposed two-step processes, with regards to the baseline *Ripper* approach. Notwithstanding, all of the three approaches follow the same trend. Notice that `fraud` is the is more problematic than the (larger) `kdd99` dataset. This is due to the complexity of the former: several attributes (mostly numerical), which clearly affect the rule generation phase.

6. Related Work

In this section, we review some seminal works from the current literature, that are most closely related to ours. Discussion aims to establish suitable connections as well as highlight major differences with respect to two major categories of approaches. We begin with an overview of research in classification within imprecise domains and then we cover some approaches to associative classification

6.1. Classification in Imprecise Domains

A wealth of approaches to learning classification models within imprecise domains exists in the literature, whose emphasis is mainly at addressing the issues related to class imbalance and different misclassification costs. We next provide an overview of some major methods, which is by no means exhaustive. The interested reader is referred to (Weiss, 2004) for a more a comprehensive survey on these topics.

Cost-sensitive learning methods (Elkan, 2001; Pazzani et al, 1994) have been explored for accounting the issues related to rare classes and different misclassification costs. The idea is to bias the learning process towards rare classes by assigning an appropriately higher value to the recognition of the minority class(es) with respect to the identification of the majority class(es). The resulting classification model has hence broader decision regions associated to the minority class(es), would boundaries are suitably extended via the specification of misclassification costs to cover more cases from the minority class(es), even if at the expense of an increased number of (misclassified) cases from the majority class(es). Nonetheless, the domain-specific information on the individual misclassification costs is seldom known or hardly quantifiable in an objective manner whenever related to domain experts' subjectiveness.

Various specific evaluation metrics have been also investigated for dealing with rare classes and different misclassification costs. The starting point here is that classification accuracy is not well-suited for imprecise domains, since it is strongly biased against rare classes and assumes equal misclassification costs. This has caused the widespread use of some alternative metrics in imprecise domains. ROC analysis is commonly used in machine learning for visualizing and evaluating the performance of classifiers. In particular, within an imprecise domain, the ROC space allows to decouple classifier performance from knowledge of both class and cost distributions. The overall performance can be summarized into a single figure, namely the Area Under the ROC Curve

(AUC), which is not biased in favor of the majority classes. The approach in (Provost and Fawcett, 2001) proposes an elegant framework that combines ROC analysis, decision analysis and computational geometry for robust classification in imprecise domains. However, a disadvantage of the method is that it requires the apriori identification of some classifiers, whose ROC curves are dominating for certain operating conditions. This clearly involves the selection and exploitation of different induction schemes to learn as many classification models under various operating conditions as well as their experimental evaluation for the purpose of identifying those areas of the ROC space, in which the curve of one classifier dominates over the others. Such a preliminary process also impacts the time efficiency of building the ROC convex hull.

Sampling involves altering the original class distribution for the purpose of attenuating or removing rarity. There are two basic forms of sampling. In particular, undersampling (Kubat and Matwin, 1997) aims at filtering cases from the majority classes, while retaining the initial population of the minority classes. Oversampling (Japkowicz, 2000) is instead devoted to replicate examples from the minority classes. Both methods have disadvantages. Precisely, undersampling wastes potentially significant examples from the majority classes that may be useful to enforce class separation, thereby hindering the performance of the resulting classifier. Oversampling prevents from missing certain portions of the data space, in which a very small number of cases from the minority classes are located. This leads to the formation of the associated decision regions. Replication clearly involves augmenting the duration of the learning process. Also, since no new information is injected into the training data, oversampling is also susceptible to overfitting especially when data is noisy. In some circumstances, this could lead to the formation of classification rules that cover one replicated case. Advanced sampling methods have also been considered. In particular, undersampling for majority classes is coupled in (Chawla et al, 2002) with a special form of oversampling for the minority classes, that creates new synthetic cases from these latter classes. The technique is effective at inducing a stronger generalization for the minority classes which neatly contrasts to the specialization induced by pure replication. However, it is still susceptible to overfitting. Progressive sampling (Weiss and Provost, 2003) approximates the best class distribution for learning by iteratively adding to some initial training data a certain proportion of cases from the majority and minority classes, by using a geometric sampling schedule. The method is empirically proven to converge towards a nearly optimal class distribution for training. Nonetheless, it assumes the existence of costs for procuring additional training data and thus it is actually useful when procurement costs are known.

Cost-sensitive boosting (Ting, 2000) has been considered for addressing two major characteristics of imprecise environments, namely the rare classes and the different misclassification costs. Boosting is an iterative meta-technique for learning ensemble classifiers, that associates a weight with each training data. Weights determine the probability with which the corresponding training cases are adaptively sampled at each iteration for the purpose of forming a new dataset. The latter is used to learn a classifier through the application of some basic learning scheme. Cost-sensitive boosting lends to being used for improving the recognition of minority classes (Fan et al, 1999; Joshi et al, 2001; Chawla et al, 2003), since the latter are more error-prone w.r.t. to majority classes and, hence, their weights are suitably increased. While weight updating is uniform in pure boosting, i.e. no focus is paid on differentiating between correct and incorrect predictions of a certain kind, cost-sensitive boosting assigns varying weights to training cases on the basis of their classifications (e.g., in the two-class scenario, TP, FP, TN and FN). The weight updating process in (Fan et al, 1999) incorporates a misclassification cost adjustment function: the weights assigned to misclassified (resp. classi-

fied) cases from a minority class are more aggressively (resp. conservatively) increased (resp. decreased) with respect to the ones associated to misclassified (resp. classified) cases from majority classes. However, since no distinction is made between cases from a minority class that are incorrectly classified into a majority class and the viceversa, the approach in (Fan et al, 1999) may overly favor recall at the expense of a much lower precision. The latter limitation is avoided in (Joshi et al, 2001) through a finer weight modification scheme. A criticism to such approaches is that cost-sensitive boosting may incur into overfitting (Weiss, 2004), by progressively increasing the weights for cases of the minority classes that are misclassified. For the purpose of avoiding overfitting and better catching the minority class, synthetic creation of cases of the minority class and boosting are combined in (Chawla et al, 2003). At each boosting iteration, a certain amount of artificial cases from the minority class are created. This allows to sample a higher number of cases from such a class, which ultimately enables the basic learning scheme to focus more on (i.e. to learn more general decision regions for) the minority class without modifying the weights of the training cases. However, it is not clear how to establish the appropriate amount of synthetic minority-class cases to generate. Besides the specific disadvantages of the enumerated methods, cost-sensitive boosting presents some general weaknesses when used for learning classification models in an imprecise domain. One such a weakness follows from the well-known inability of boosting at properly working in the presence of noise. Additionally, there is not general guarantee that it can improve the recognition of the rare class(es) since its performance is strictly dependent on the performance of the basic learning scheme. If the underlying scheme always achieves low recall or precision on the rare class(es) of an imprecise domains, the performance of boosting is also poor (Joshi et al, 2002).

Finally, segmentation (Weiss, 2004) is another major method for catching rare classes in imprecise domains. The underlying idea is to suitably divide the data space into disjoint regions, wherein globally rare classes tend to become less rare. Within each such a region there are two possibilities. The density of rarity is (much) higher w.r.t. the density in the whole training data. This clearly allows to focus on the rarities local to the region, which are also less affected by noise. Alternatively, the density of rarity is (much) lower than the corresponding density in the training data. In this circumstance, rarity becomes nearly unidentifiable in the specific region. Nonetheless, this is acceptable in practice, since most of the original class rarity is still captured within other regions. Segmentation is adopted in subsection 3.2 to better model rare classes. This is achieved by training a probabilistic generative model over the subset of training data covered by each individual rule of an associative classifier.

6.2. Associative Classification

Several approaches to associative classification have been proposed in the literature, with differences in three major aspects, i.e the discovery of class association rules, the extraction of a classifier and the class prediction for unlabeled cases (Thabtah, 2007). The search for class association rules is a critical aspect due to the implied amount of computation. A greedy strategy that refines the basic FOIL algorithm (Quinlan and Cameron-Jones, 1993) is leveraged in (Xin and Han, 2003). Search strategies based on Apriori (Agrawal and Srikant, 1994) are applied in (Liu et al, 1998; Liu et al, 2000; Antonie and Zaïane, 2002), whereas a variant of the FP-growth algorithm (Han and Yin, 2000) is at the basis of (Li et al, 2001) A row enumeration method (Cong et al, 2004) is used in (Arunasalam and Chawla, 2006). The covering rules with highest confidence for each training case are directly mined in (Wang and Karypis, 2005). In many cases, the

huge number of resulting class association rules, that may potentially overfit the training data, is pruned to obtain a compact associative classifier. For instance, redundant rules, i.e. those rules whose confidence is lower than the confidence of more general rules, are pruned in (Li et al, 2001; Antonie and Zaïane, 2002). χ^2 testing is performed in (Li et al, 2001) to filter those rules whose antecedents and consequents are not positively correlated. The minimum class support and the complement class support are introduced in (Liu et al, 2000; Arunasalam and Chawla, 2006), respectively, to prune those rules whose support or confidence is lower that than a threshold automatically identified for the targeted class. The database coverage method is used in (Liu et al, 1998; Liu et al, 2000; Antonie and Zaïane, 2002) to extract a classifier from the class association rules. As far as the prediction of an unlabeled case is concerned, associative classification methods can be divided into two categories. On one hand are those approaches that exploit the top-quality rule covering the case (Liu et al, 1998; Liu et al, 2000). On the other hand, the eventual prediction is delivered by taking into account multiple rules applicable to the case (Li et al, 2001; Xin and Han, 2003; Wang and Karypis, 2005; Antonie and Zaïane, 2002).

Learning in imprecise domains has not been a primary focus of research on associative classification, apart from few approaches such as (Liu et al, 1998; Liu et al, 2000), that account for imbalance in class distribution. The MINECARS scheme discussed in subsection 3.1 situates in the foregoing framework of related works as follows. It is an Apriori-based search method designed for mining class association rules within imprecise domains. In principle, MINECARS can exhaustively search for class association rules and, hence, is potentially capable to unveil meaningful associations among rare items. Therein, one inherent difficulty is that catching such associations may require a very low value for the minimum support threshold. This would ultimately make the search for meaningful class association rules computationally intractable, because of the resulting combinatorial explosion in the number of ways in which the individual items would be associated with one another. In order to ensure tractability, minimum class support (Liu et al, 2000) and complement class support (Arunasalam and Chawla, 2006) are both integrated into MINECARS. The former requires setting one parameter, a global threshold for minimum class support, that is dynamically adjusted to become a class-specific minimum support threshold. Such a mechanism enables the discovery of an appropriate number of rules within each individual class, thus avoiding that the rules targeting the majority classes overwhelm the ones predicting the minority classes. Additionally, complement class support is used to guarantee that the discovered class association rules are positively correlated, without performing any further correlation analysis and testing (which are instead an essential requirement in approaches such as (Li et al, 2001)). Complement class support avoids the flaw with the traditional support-and-confidence framework, wherein the antecedents and consequents of high confidence rules may be negatively correlated in the presence of imbalanced class distributions. Within imprecise domains, such a flaw is clearly a concern for primary approaches such as (Liu et al, 2000; Wang and Karypis, 2005). The potentially large set of class association rules found by MINECARS is pruned through a suitable overfitting avoidance strategy, that removes whole rules on the basis of statistical arguments. As to the classification of unlabeled cases, two schemes are adopted in subsection 3.2 for class prediction that consider multiple class association rules as well as their corresponding probabilistic generative models.

7. Conclusions and Future Work

This manuscript proposed two probabilistic frameworks for improving the performance of rule-based classification in highly-imprecise (multi-class) learning environments.

In particular, the *global-to-local* scheme couples the individual rules of an associative classifier with as many local probabilistic discriminative models. The individual model is built over the coverage of each classifier rule and is, then, involved into the classification process for more effectively dealing with those globally rare classes, that are likely to become less rare within the coverage. Two novel schemes for a tight integration between associative and local probabilistic models were discussed.

Instead, the *local-to-global* scheme elegantly and seamlessly integrates associative classification with discriminative learning through the maximum entropy framework in order to boost the overall classification as a side effect of mutual influence.

A massive evaluation revealed that both learning frameworks are competitive and often superior in classification performance w.r.t. established rule-based competitors.

The ongoing research efforts are mainly geared towards the improvement of the accuracy of the local probabilistic models in the *global-to-local* scheme through the analysis of ROC curves. The point is that the classification threshold typically used in our framework assigns a class label when the associated probability is higher than 0.5. However, the latter may not necessarily be the best threshold, especially if we consider the bias introduced by the CAR associated with the probabilistic classifier. In general, lower thresholds produce improvements in recall, by contemporarily degrading precision as a side effect. However, as suggested by figure 5 where a better threshold value can be obtained in correspondence to the $(0.8, 0.01)$ pair (corresponding to the threshold 0.2), by automatically choosing the best class-specific threshold, probabilistic smoothing can still allow to remove some locality effects within the CAR and maintain high precision as well.

A. Further Details Classification Performances

The following tables 5, 6, 7 and 8 report the values of, respectively, precision, recall, f-measure and AUC achieved by both the devised techniques and their competitors over each class of the chosen datasets. Precisely, each row of such tables indicates the corresponding classification performance over a particular class of a certain dataset, whose identities are specified in the respective entry of the **Dataset and class** column according to the notation `dataset_class`. Moreover, for each selected dataset, the above tables also include an additional summarization row, that specifies the average classification performance across all classes of that dataset. Such rows are distinguished by the values in the respective entries of the **Dataset and class** column, which read as `mean(dataset)`. Notice that, within every table, the best classification performance over each class of the chosen datasets is highlighted in bold.

References

Agrawal R, Srikant R (1994) Fast Algorithms for Mining Association Rules. In: Proceedings of International Conference on Very Large Data Bases, 1994, pp 487–499

Antonie ML, Zaïane OR (2002) Text Document Categorization by Term Association. In: Proceedings of IEEE International Conference on Data Mining, 2002, pp 19–26

Table 5. Precision

Dataset and class	Size	(1)	(2)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
anneal_1	8	**1.000**	0.800	0.778	0.778	**1.000**	**1.000**	**1.000**	**1.000**	0.013
anneal_2	99	0.961	0.971	0.960	0.960	0.980	0.980	0.961	0.961	**1.000**
anneal_3	684	0.988	0.983	0.990	0.990	0.997	0.997	0.988	0.988	**1.000**
anneal_5	67	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
anneal_U	40	0.947	0.941	0.900	0.900	**1.000**	**1.000**	0.947	0.947	**1.000**
mean(anneal)	898	0.984	0.979	0.981	0.981	**0.996**	**0.996**	0.984	0.984	**0.996**
balance-scale_L	288	0.813	0.603	0.832	0.826	0.859	**0.864**	0.810	0.813	0.567
balance-scale_B	49	0.000	**0.163**	0.083	0.000	0.000	0.000	0.000	0.000	0.000
balance-scale_R	288	0.803	0.624	0.826	0.826	0.878	0.851	0.806	0.803	**0.918**
mean(balance-scale)	625	0.745	0.578	0.770	0.761	**0.800**	0.790	0.745	0.745	0.684
breast-cancer_no-rec.	201	**0.761**	0.745	0.749	0.749	0.760	0.760	0.752	0.752	0.751
breast-cancer_rec.	85	0.583	0.535	0.475	0.475	0.547	0.527	0.568	0.568	**0.905**
mean(breast-cancer)	286	0.708	0.682	0.667	0.667	0.697	0.691	0.697	0.697	**0.797**
horse-colic_yes	232	**0.854**	0.849	0.851	0.851	0.852	0.852	0.840	0.840	0.819
horse-colic_no	136	0.820	0.845	0.759	0.759	0.765	0.765	0.776	0.776	**0.945**
mean(horse-colic)	368	0.841	0.848	0.817	0.817	0.820	0.820	0.816	0.816	**0.866**
credit-rating_+	307	0.828	0.832	0.830	0.830	0.792	0.792	**0.863**	**0.863**	0.787
credit-rating_-	383	0.884	0.871	0.770	0.770	0.835	0.835	0.875	0.875	**0.922**
mean(credit-rating)	690	0.859	0.854	0.797	0.797	0.816	0.816	**0.869**	**0.869**	0.862
german_credit_good	700	0.759	0.779	**0.803**	**0.803**	0.778	0.778	0.790	0.790	0.700
german_credit_bad	300	0.544	0.504	**0.599**	**0.599**	0.495	0.495	**0.599**	**0.599**	0.250
mean(german_credit)	1000	0.694	0.696	0.742	0.742	0.693	0.693	**0.733**	**0.733**	0.565
diabetes_negative	500	0.793	0.790	0.776	0.776	0.758	0.758	**0.798**	**0.798**	0.648
diabetes_positive	268	**0.684**	0.667	0.629	0.629	0.570	0.570	**0.684**	**0.684**	0.000
mean(diabetes)	768	0.755	0.747	0.725	0.725	0.692	0.692	**0.758**	**0.758**	0.422
Glass_build wind float	70	**0.723**	0.667	0.530	0.535	0.721	0.721	**0.723**	**0.723**	0.516
Glass_build wind non-float	76	0.652	0.640	0.625	0.625	0.684	0.675	0.656	0.656	**0.939**
Glass_vehic wind float	17	0.250	0.545	0.188	0.200	0.267	0.267	0.250	0.250	**1.000**
Glass_containers	13	0.667	0.750	0.769	0.769	0.667	0.643	0.667	0.714	**1.000**
Glass_tableware	9	0.700	0.727	0.700	0.700	0.700	0.700	0.700	0.700	**1.000**
Glass_headlamps	29	**0.852**	**0.852**	**0.852**	**0.852**	0.815	0.815	0.846	**0.852**	0.813
mean(Glass)	214	0.673	0.680	0.602	0.604	0.680	0.676	0.674	0.677	**0.795**
cleveland-heart_<50	165	0.834	0.853	0.871	0.871	0.801	0.801	0.838	0.838	**0.910**
cleveland-heart_>50_1	138	0.793	0.782	0.797	0.797	0.766	0.766	0.816	0.816	**0.898**
mean(cleveland-heart)	303	0.815	0.821	0.837	0.837	0.785	0.785	0.828	0.828	**0.904**
hungarian-heart_<50	188	0.808	0.820	0.862	0.862	0.815	0.815	0.858	0.858	**0.866**
hungarian-heart_>50_1	106	0.736	0.775	0.762	0.762	0.676	0.676	0.804	0.804	**0.859**
mean(hungarian-heart)	294	0.782	0.804	0.826	0.826	0.765	0.765	0.838	0.838	**0.864**
heart-statlog_absent	150	0.789	0.757	0.805	0.805	0.776	0.776	0.790	0.790	**0.870**
heart-statlog_present	120	0.789	0.703	0.752	0.752	0.707	0.707	0.796	0.796	**0.917**
mean(heart-statlog)	270	0.789	0.733	0.782	0.782	0.745	0.745	0.793	0.793	**0.891**
hepatitis_DIE	32	0.458	0.618	0.458	0.458	**0.533**	**0.533**	0.529	0.529	0.516
hepatitis_LIVE	123	0.840	0.909	0.840	0.840	0.872	0.872	0.884	0.884	**1.000**
mean(hepatitis)	155	0.761	0.849	0.761	0.761	0.802	0.802	0.811	0.811	**0.900**
ionosphere_b	126	0.852	0.922	0.761	0.761	**0.950**	**0.950**	0.858	0.858	0.592
ionosphere_g	225	0.904	0.915	0.901	0.901	0.880	0.880	0.900	0.900	**1.000**
mean(ionosphere)	351	0.885	**0.918**	0.851	0.851	0.905	0.905	0.885	0.885	0.853
labor_bad	20	0.684	0.700	0.684	0.684	0.684	0.684	0.722	0.722	**1.000**
labor_good	37	0.816	0.838	0.816	0.816	0.816	0.816	0.821	0.821	**1.000**
mean(labor)	57	0.770	0.789	0.770	0.770	0.770	0.770	0.786	0.786	**1.000**
lymphography_normal	2	0.000	0.000	0.000	0.000	0.000	0.000	**1.000**	**1.000**	0.000
lymphography_metastases	81	0.798	0.775	0.825	0.825	0.807	0.807	0.807	0.807	**1.000**
lymphography_malign	61	0.750	0.788	0.719	0.719	0.742	0.742	0.738	0.750	**1.000**
lymphography_fibrosis	4	0.750	0.600	0.750	0.750	**1.000**	**1.000**	**1.000**	0.750	**1.000**
mean(lymphography)	148	0.766	0.765	0.768	0.768	0.775	0.775	0.786	0.785	**1.000**
sick_negative	3541	0.992	0.993	0.990	0.990	**0.991**	**0.991**	**0.991**	**0.991**	0.977
sick_sick	231	0.852	0.887	0.859	0.859	0.896	0.896	0.888	0.888	**0.973**
mean(sick)	3772	0.983	**0.986**	0.982	0.982	0.985	0.985	0.985	0.985	0.976
sonar_Rock	97	0.725	0.775	0.694	0.694	0.795	0.795	0.744	0.744	**0.933**
sonar_Mine	111	0.735	0.830	0.780	0.780	0.775	0.775	0.730	0.730	**1.000**
mean(sonar)	208	0.730	0.804	0.740	0.740	0.785	0.785	0.736	0.736	**0.969**
fraud_0	22109	**0.980**	0.977	0.974	0.974	0.970	0.970	0.977	0.977	0.972
fraud_1	11667	0.895	0.889	0.401	0.401	0.891	0.891	**0.899**	**0.899**	0.893
fraud_2	8167	**0.879**	0.870	0.550	0.550	0.849	0.849	0.875	0.876	0.851
fraud_3	3499	0.878	0.874	0.767	0.767	0.877	0.877	**0.895**	**0.895**	0.880
mean(fraud)	45442	**0.932**	0.927	0.734	0.734	0.921	0.921	**0.932**	**0.932**	0.924
kdd99_r2l	900	0.812	0.828	0.787	0.798	**0.832**	0.829	0.792	0.812	0.783
kdd99_u2r	40	0.403	**0.532**	0.325	0.325	0.333	0.392	0.403	0.403	0.253
kdd99_dos	366458	0.999	0.999	0.992	0.992	**1.000**	**1.000**	**1.000**	**1.000**	0.974
kdd99_probe	3707	0.992	0.976	0.992	0.992	**0.995**	**0.995**	0.992	0.992	0.966
kdd99_normal	92270	0.998	0.998	**0.999**	**0.999**	**0.999**	**0.999**	0.998	0.998	0.742
mean(kdd99)	464075	**0.999**	**0.999**	0.993	0.993	**0.999**	**0.999**	**0.999**	**0.999**	0.972

Table 6. Recall

Dataset and class	Size	(1)	(2)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
anneal_1	8	0.500	0.500	0.875	0.875	0.875	0.875	0.500	0.500	**1.000**
anneal_2	99	**1.000**	**1.000**	0.980	0.980	**1.000**	**1.000**	**1.000**	**1.000**	0.758
anneal_3	684	0.991	0.991	0.985	0.985	**0.997**	**0.997**	0.991	0.991	0.637
anneal_5	67	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
anneal_U	40	0.900	0.800	0.900	0.900	**0.975**	**0.975**	0.900	0.900	0.425
mean(anneal)	898	0.984	0.980	0.981	0.981	**0.996**	**0.996**	0.984	0.984	0.670
balance-scale_L	288	0.889	0.608	0.892	0.892	0.913	0.903	0.889	0.889	**0.990**
balance-scale_B	49	0.000	**0.156**	0.020	0.000	0.000	0.000	0.000	0.000	0.000
balance-scale_R	288	0.865	0.624	0.872	0.872	0.872	**0.892**	0.865	0.865	0.389
mean(balance-scale)	625	0.808	0.580	0.814	0.813	0.822	**0.827**	0.808	0.808	0.635
breast-cancer_no-rec.	201	0.900	0.900	0.846	0.846	0.881	0.866	0.905	0.905	**0.990**
breast-cancer_rec	85	0.329	0.271	0.329	0.329	0.341	**0.354**	0.294	0.294	0.224
mean(breast-cancer)	286	0.731	0.713	0.692	0.692	0.720	0.714	0.724	0.724	**0.762**
horse-colic_yes	232	0.905	0.922	0.862	0.862	0.866	0.866	0.879	0.879	**0.978**
horse-colic_no	136	0.735	0.721	**0.743**	**0.743**	**0.743**	**0.743**	0.713	0.713	0.632
mean(horse-colic)	368	0.842	0.848	0.818	0.818	0.821	0.821	0.818	0.818	**0.851**
credit-rating_+	307	0.860	0.840	0.668	0.668	0.795	0.795	0.840	0.840	**0.915**
credit-rating_-	383	0.856	0.864	0.890	0.890	0.833	0.833	**0.893**	**0.893**	0.802
mean(credit-rating)	690	0.858	0.854	0.791	0.791	0.816	0.816	**0.870**	**0.870**	0.852
german_credit_good	700	0.873	0.801	0.853	0.853	0.793	0.793	0.867	0.867	**0.996**
german_credit_bad	300	0.353	0.470	**0.513**	**0.513**	0.473	0.473	0.463	0.463	0.003
mean(german_credit)	1000	0.717	0.702	**0.751**	**0.751**	0.697	0.697	0.746	0.746	0.698
diabetes_negative	500	0.856	0.844	0.824	0.824	0.784	0.784	0.852	0.852	**0.988**
diabetes_positive	268	0.582	0.582	0.556	0.556	0.534	0.534	**0.597**	**0.597**	0.000
mean(pima_diabetes)	768	0.760	0.753	0.730	0.730	0.697	0.697	**0.763**	**0.763**	0.643
Glass_build wind float	70	0.671	0.743	0.757	0.771	0.700	0.700	0.671	0.671	**0.943**
Glass_build wind non-float	76	0.763	0.632	0.395	0.395	0.711	0.711	**0.776**	**0.776**	0.408
Glass_vehic wind float	17	0.118	**0.353**	0.176	0.176	0.235	0.235	0.118	0.118	0.118
Glass_containers	13	0.769	0.692	0.769	0.769	0.769	0.692	0.769	0.769	**0.923**
Glass_tableware	9	0.778	**0.889**	0.778	0.778	0.778	0.778	0.778	0.778	0.778
Glass_headlamps	29	0.793	0.793	0.793	0.793	0.759	0.759	0.759	0.793	**0.897**
mean(Glass)	214	0.687	0.682	0.589	0.593	0.682	0.678	0.687	**0.692**	0.673
cleveland-heart_<50	165	0.824	0.806	0.818	0.818	0.806	0.806	0.848	0.848	**0.915**
cleveland-heart_>50_1	138	0.804	0.833	0.855	0.855	0.761	0.761	0.804	0.804	**0.891**
mean(cleveland-heart	303	0.815	0.818	0.835	0.835	0.785	0.785	0.828	0.828	**0.904**
hungarian-heart_<50	188	0.872	0.894	0.867	0.867	0.819	0.819	0.899	0.899	**0.931**
hungarian-heart_>50_1	106	0.632	0.651	**0.755**	**0.755**	0.670	0.670	0.736	0.736	0.745
mean(hungarian-heart	294	0.786	0.806	0.827	0.827	0.765	0.765	0.840	0.840	**0.864**
heart-statlog_absent	150	0.847	0.767	0.800	0.800	0.760	0.760	0.853	0.853	**0.940**
heart-statlog_present	120	0.717	0.692	0.758	0.758	0.725	0.725	0.717	0.717	**0.825**
mean(heart-statlog)	270	0.789	0.733	0.781	0.781	0.744	0.744	0.793	0.793	**0.889**
hepatitis_DIE	32	0.344	0.656	0.344	0.344	0.500	0.500	0.563	0.563	**1.000**
hepatitis_LIVE	123	**0.894**	**0.894**	**0.894**	**0.894**	0.886	0.886	0.870	0.870	0.756
mean(hepatitis)	155	0.781	**0.845**	0.781	0.781	0.806	0.806	0.806	0.806	0.806
ionosphere_b	126	0.825	0.841	0.833	0.833	0.762	0.762	0.817	0.817	**1.000**
ionosphere_g	225	0.920	0.960	0.853	0.853	**0.978**	**0.978**	0.924	0.924	0.613
mean(ionosphere)	351	0.886	**0.917**	0.846	0.846	0.900	0.900	0.886	0.886	0.752
labor_bad	20	0.650	0.700	0.650	0.650	0.650	0.650	0.650	0.650	**1.000**
labor_good	37	0.838	0.838	0.838	0.838	0.838	0.838	0.865	0.865	**1.000**
mean(labor)	57	0.772	0.789	0.772	0.772	0.772	0.772	0.789	0.789	**1.000**
lymphography_normal	2	0.000	0.000	0.000	0.000	0.000	0.000	**0.500**	**0.500**	0.000
lymphography_metastases	81	0.827	0.852	0.815	0.815	0.827	0.827	0.827	0.827	**0.901**
lymphography_malign	61	0.738	0.672	0.754	0.754	0.754	0.754	0.738	0.738	**0.885**
lymphography_fibrosis	4	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	**1.000**
mean(lymphography)	148	0.777	0.764	0.777	0.777	0.784	0.784	0.784	0.784	**0.897**
sick_negative	3541	0.990	0.993	0.991	0.991	0.994	0.994	0.993	0.993	**0.999**
sick_sick	231	0.870	**0.887**	0.844	0.844	0.857	0.857	0.861	0.861	0.632
mean(sick)	3772	0.983	**0.986**	0.982	0.982	0.985	0.985	0.985	0.985	0.976
sonar_Rock	97	0.680	0.814	0.773	0.773	0.722	0.722	0.660	0.660	**1.000**
sonar_Mine	111	0.775	0.793	0.703	0.703	0.838	0.838	0.802	0.802	**0.937**
mean(sonar)	208	0.731	0.803	0.736	0.736	0.784	0.784	0.736	0.736	**0.966**
fraud_0	22109	0.983	0.975	0.886	0.886	0.985	0.985	**0.988**	**0.988**	0.985
fraud_1	11667	**0.904**	0.898	0.850	0.850	0.880	0.880	0.899	0.899	0.885
fraud_2	8167	0.838	**0.860**	0.829	0.829	0.834	0.834	0.843	0.843	0.839
fraud_3	3499	**0.924**	0.879	0.828	0.828	0.870	0.870	0.911	0.912	0.873
mean(fraud)	45442	0.932	0.927	0.869	0.869	0.922	0.922	**0.933**	**0.933**	0.925
kdd99_r2l	900	0.941	0.936	0.946	0.946	0.950	**0.951**	0.941	0.941	0.944
kdd99_u2r	40	**0.625**	**0.625**	**0.625**	**0.625**	**0.625**	0.500	**0.625**	**0.625**	0.600
kdd99_dos	366458	0.999	0.999	0.999	0.999	**1.000**	**1.000**	**1.000**	**1.000**	0.920
kdd99_probe	3707	**0.982**	0.979	0.978	**0.982**	0.980	0.980	0.981	**0.982**	0.980
kdd99_normal	92270	**0.997**	**0.997**	0.964	0.964	**0.997**	**0.997**	**0.997**	**0.997**	0.899
mean(kdd99)	464075	**0.999**	**0.999**	0.992	0.992	**0.999**	**0.999**	**0.999**	**0.999**	0.918

Table 7. F-Measure

Dataset and class	Size	(1)	(2)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
anneal_1	8	0.667	0.615	0.824	0.824	**0.933**	**0.933**	0.667	0.667	0.026
anneal_2	99	0.980	0.985	0.970	0.970	**0.990**	**0.990**	0.980	0.980	0.862
anneal_3	684	0.990	0.987	0.988	0.988	**0.997**	**0.997**	0.990	0.990	0.779
anneal_5	67	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
anneal_U	40	0.923	0.865	0.900	0.900	**0.987**	**0.987**	0.923	0.923	0.596
mean(anneal)	898	0.984	0.979	0.981	0.981	**0.996**	**0.996**	0.984	0.984	0.793
balance-scale_L	288	0.849	0.606	0.861	0.858	**0.886**	0.883	0.848	0.849	0.721
balance-scale_B	49	0.000	**0.159**	0.033	0.000	0.000	0.000	0.000	0.000	0.000
balance-scale_R	288	0.833	0.624	0.848	0.848	**0.875**	0.871	0.834	0.833	0.546
mean(balance-scale)	625	0.775	0.579	0.790	0.786	**0.811**	0.808	0.775	0.775	0.584
breast-cancer_no-rec.	201	0.825	0.815	0.794	0.794	0.816	0.810	0.822	0.822	**0.854**
breast-cancer_rec	85	0.421	0.359	0.389	0.389	0.420	**0.423**	0.388	0.388	0.358
mean(breast-cancer)	286	0.705	0.680	0.674	0.674	0.698	0.695	0.693	0.693	**0.707**
horse-colic_yes	232	0.879	0.884	0.857	0.857	0.859	0.859	0.859	0.859	**0.892**
horse-colic_no	136	0.775	**0.778**	0.751	0.751	0.754	0.754	0.743	0.743	0.758
mean(horse-colic)	368	0.840	**0.845**	0.818	0.818	0.820	0.820	0.816	0.816	0.842
credit-rating_+	307	0.843	0.836	0.740	0.740	0.793	0.793	**0.851**	**0.851**	0.846
credit-rating_-	383	0.870	0.868	0.826	0.826	0.834	0.834	**0.884**	**0.884**	0.858
mean(credit-rating)	690	0.858	0.854	0.788	0.788	0.816	0.816	**0.869**	**0.869**	0.853
german_credit_good	700	0.812	0.790	**0.827**	**0.827**	0.786	0.786	**0.827**	**0.827**	0.822
german_credit_bad	300	0.428	0.486	**0.553**	**0.553**	0.484	0.484	0.523	0.523	0.007
mean(german_credit)	1000	0.697	0.699	**0.745**	**0.745**	0.695	0.695	0.736	0.736	0.577
diabetes_negative	500	0.823	0.816	0.799	0.799	0.771	0.771	**0.824**	**0.824**	0.783
diabetes_positive	268	0.629	0.622	0.590	0.590	0.551	0.551	**0.637**	**0.637**	0.000
mean(pima_diabetes)	768	0.755	0.748	0.726	0.726	0.694	0.694	**0.759**	**0.759**	0.510
Glass_build wind float	70	0.696	0.703	0.624	0.632	**0.710**	**0.710**	0.696	0.696	0.667
Glass_build wind non-float	76	0.703	0.636	0.484	0.484	0.697	0.692	**0.711**	**0.711**	0.569
Glass_vehic wind float	17	0.160	**0.429**	0.182	0.188	0.250	0.250	0.160	0.160	0.211
Glass_containers	13	0.714	0.720	0.769	0.769	0.714	0.667	0.714	0.741	**0.960**
Glass_tableware	9	0.737	0.800	0.737	0.737	0.737	0.737	0.737	0.737	**0.875**
Glass_headlamps	29	0.821	0.821	0.821	0.821	0.786	0.786	0.800	0.821	**0.852**
mean(Glass)	214	0.676	0.678	0.579	0.582	**0.680**	0.676	0.676	**0.680**	0.647
cleveland-heart_<50	165	0.829	0.829	0.844	0.844	0.804	0.804	0.843	0.843	**0.912**
cleveland-heart_>50_1	138	0.799	0.807	0.825	0.825	0.764	0.764	0.810	0.810	**0.895**
mean(cleveland-heart	303	0.815	0.819	0.835	0.835	0.785	0.785	0.828	0.828	**0.904**
hungarian-heart_<50	188	0.839	0.855	0.865	0.865	0.817	0.817	0.878	0.878	**0.897**
hungarian-heart_>50_1	106	0.680	0.708	0.758	0.758	0.673	0.673	0.768	0.768	**0.798**
mean(hungarian-heart	294	0.782	0.802	0.826	0.826	0.765	0.765	0.838	0.838	**0.862**
heart-statlog_absent	150	0.817	0.762	0.803	0.803	0.768	0.768	0.821	0.821	**0.904**
heart-statlog_present	120	0.751	0.697	0.755	0.755	0.716	0.716	0.754	0.754	**0.868**
mean(heart-statlog)	270	0.788	0.733	0.782	0.782	0.745	0.745	0.791	0.791	**0.888**
hepatitis_DIE	32	0.393	0.636	0.393	0.393	0.516	0.516	0.545	0.545	**0.681**
hepatitis_LIVE	123	0.866	**0.902**	0.866	0.866	0.879	0.879	0.877	0.877	0.861
mean(hepatitis)	155	0.768	**0.847**	0.768	0.768	0.804	0.804	0.809	0.809	0.824
ionosphere_b	126	0.839	**0.880**	0.795	0.795	0.846	0.846	0.837	0.837	0.743
ionosphere_g	225	0.912	**0.937**	0.877	0.877	0.926	0.926	0.912	0.912	0.760
mean(ionosphere)	351	0.886	**0.916**	0.848	0.848	0.897	0.897	0.885	0.885	0.754
labor_bad	20	0.667	0.700	0.667	0.667	0.667	0.667	0.684	0.684	**1.000**
labor_good	37	0.827	0.838	0.827	0.827	0.827	0.827	0.842	0.842	**1.000**
mean(labor)	57	0.771	0.789	0.771	0.771	0.771	0.771	0.787	0.787	**1.000**
lymphography_normal	2	0.000	0.000	0.000	0.000	0.000	0.000	**0.667**	**0.667**	0.000
lymphography_metastases	81	0.812	0.812	0.820	0.820	0.817	0.817	0.817	0.817	**0.948**
lymphography_malign	61	0.744	0.726	0.736	0.736	0.748	0.748	0.738	0.744	**0.939**
lymphography_fibrosis	4	0.750	0.667	0.750	0.750	0.857	0.857	0.857	0.750	**1.000**
mean(lymphography)	148	0.771	0.761	0.772	0.772	0.779	0.779	0.783	0.783	**0.946**
sick_negative	3541	0.991	**0.993**	0.990	0.990	0.992	0.992	0.992	0.992	0.988
sick_sick	231	0.861	**0.887**	0.852	0.852	0.876	0.876	0.875	0.875	0.766
mean(sick)	3772	0.983	**0.986**	0.982	0.982	0.985	0.985	0.985	0.985	0.974
sonar_Rock	97	0.702	0.794	0.732	0.732	0.757	0.757	0.699	0.699	**0.965**
sonar_Mine	111	0.754	0.811	0.739	0.739	0.805	0.805	0.764	0.764	**0.967**
mean(sonar)	208	0.730	0.803	0.736	0.736	0.783	0.783	0.734	0.734	**0.966**
fraud_0	22109	0.981	0.976	0.928	0.928	0.977	0.977	**0.982**	**0.982**	0.979
fraud_1	11667	**0.900**	0.893	0.545	0.545	0.885	0.885	0.899	0.899	0.886
fraud_2	8167	0.858	**0.865**	0.661	0.661	0.841	0.841	0.859	0.859	0.847
fraud_3	3499	0.900	0.877	0.796	0.796	0.873	0.873	**0.903**	**0.903**	0.885
mean(fraud)	45442	0.932	0.927	0.796	0.796	0.921	0.921	0.932	**0.933**	0.928
kdd99_r2l	900	0.872	0.878	0.859	0.865	**0.887**	0.886	0.860	0.872	0.856
kdd99_u2r	40	0.490	**0.575**	0.427	0.427	0.435	0.440	0.490	0.490	0.356
kdd99_dos	366458	0.999	0.999	0.996	0.996	**1.000**	**1.000**	**1.000**	**1.000**	0.946
kdd99_probe	3707	0.987	0.978	0.985	0.987	**0.988**	**0.988**	0.986	0.987	0.973
kdd99_normal	92270	0.997	0.997	0.981	0.981	**0.998**	**0.998**	0.997	0.997	0.813
mean(kdd99)	464075	**0.999**	**0.999**	0.992	0.992	**0.999**	**0.999**	**0.999**	**0.999**	0.944

Table 8. AUC per Class

Dataset and class	Size	(1)	(2)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
anneal_1	8	0.823	0.866	0.933	0.933	0.938	0.938	0.909	0.909	**0.979**
anneal_2	99	**0.999**	**0.999**	0.998	0.998	**0.999**	**0.999**	**0.999**	**0.999**	0.959
anneal_3	684	0.981	0.988	0.984	0.984	**0.997**	**0.997**	0.989	0.989	0.959
anneal_5	67	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
anneal_U	40	0.968	0.984	0.965	0.965	**0.999**	**0.999**	0.977	0.977	0.909
mean(anneal)	898	0.983	0.989	0.986	0.986	**0.997**	**0.997**	0.990	0.990	0.960
balance-scale_L	288	0.887	0.631	0.920	0.917	**0.943**	**0.943**	0.902	0.902	0.673
balance-scale_B	49	0.578	0.482	**0.664**	0.648	0.495	0.493	0.641	0.621	0.528
balance-scale_R	288	0.877	0.656	0.913	0.906	**0.935**	0.934	0.886	0.886	0.682
mean(balance-scale)	625	0.858	0.631	0.897	0.891	**0.904**	0.903	0.874	0.873	0.666
breast-cancer_no-rec.	201	0.598	0.586	0.694	0.694	0.633	0.624	0.712	0.712	**0.749**
breast-cancer_rec	85	0.598	0.586	0.694	0.694	0.633	0.625	0.712	0.712	**0.749**
mean(breast-cancer)	286	0.598	0.586	0.694	0.694	0.633	0.625	0.712	0.712	**0.749**
horse-colic_yes	232	0.823	0.863	0.821	0.821	0.803	0.803	0.871	0.871	**0.952**
horse-colic_no	136	0.823	0.863	0.821	0.821	0.803	0.803	0.871	0.871	**0.952**
mean(horse-colic)	368	0.823	0.863	0.821	0.821	0.803	0.803	0.871	0.871	**0.952**
credit-rating_+	307	0.874	0.892	0.878	0.878	0.828	0.828	0.921	0.921	**0.937**
credit-rating_-	383	0.874	0.892	0.877	0.877	0.828	0.828	0.921	0.921	**0.938**
mean(credit-rating)	690	0.874	0.892	0.878	0.878	0.828	0.828	0.921	0.921	**0.938**
german_credit_good	700	0.593	0.658	0.773	0.773	0.656	0.656	**0.774**	**0.774**	0.715
german_credit_bad	300	0.593	0.658	0.773	0.773	0.656	0.656	**0.774**	**0.774**	0.719
mean(german_credit)	1000	0.593	0.658	0.773	0.773	0.656	0.656	**0.774**	**0.774**	0.716
diabetes_negative	500	0.739	0.794	0.777	0.777	0.671	0.671	**0.807**	**0.807**	0.754
diabetes_positive	268	0.739	0.794	0.777	0.777	0.671	0.671	**0.807**	**0.807**	0.755
mean(pima_diabetes)	768	0.739	0.794	0.777	0.777	0.671	0.671	**0.807**	**0.807**	0.754
Glass_build wind float	70	0.814	0.777	0.796	0.795	0.760	0.737	**0.816**	0.798	0.761
Glass_build wind non-float	76	0.765	0.724	0.767	0.767	0.756	0.755	0.792	0.800	**0.802**
Glass_vehic wind float	17	**0.748**	0.706	0.736	0.736	0.554	0.561	0.699	0.705	0.747
Glass_containers	13	0.853	0.910	0.874	0.874	0.881	0.826	0.870	0.873	**1.000**
Glass_tableware	9	0.883	**0.937**	0.882	0.882	0.931	0.935	0.930	0.933	0.907
Glass_headlamps	29	0.859	0.892	0.875	0.875	0.876	0.880	0.865	0.867	**0.960**
mean(Glass)	214	0.803	0.783	0.800	0.800	0.772	0.763	0.813	0.811	**0.822**
cleveland-heart_<50	165	0.831	0.829	0.870	0.870	0.806	0.806	0.898	0.898	**0.971**
cleveland-heart_>50_1	138	0.831	0.829	0.870	0.870	0.806	0.806	0.898	0.898	**0.973**
mean(cleveland-heart	303	0.831	0.829	0.870	0.870	0.806	0.806	0.898	0.898	**0.972**
hungarian-heart_<50	188	0.730	0.859	0.891	0.891	0.794	0.794	0.898	0.898	**0.920**
hungarian-heart_>50_1	106	0.730	0.859	0.891	0.891	0.794	0.794	0.898	0.898	**0.922**
mean(hungarian-heart	294	0.730	0.859	0.891	0.891	0.794	0.794	0.898	0.898	**0.920**
heart-statlog_absent	150	0.781	0.736	0.803	0.803	0.752	0.752	0.796	0.796	**0.961**
heart-statlog_present	120	0.781	0.736	0.755	0.755	0.752	0.752	0.796	0.796	**0.961**
mean(heart-statlog)	270	0.781	0.736	0.782	0.782	0.752	0.752	0.796	0.796	**0.961**
hepatitis_DIE	32	0.664	0.828	0.667	0.667	0.715	0.715	0.838	0.838	**0.992**
hepatitis_LIVE	123	0.664	0.828	0.667	0.667	0.715	0.715	0.838	0.838	**0.992**
mean(hepatitis)	155	0.664	0.828	0.667	0.667	0.715	0.715	0.838	0.838	**0.992**
ionosphere_b	126	0.896	0.916	0.893	0.893	0.908	0.908	0.900	0.900	**0.966**
ionosphere_g	225	0.896	0.916	0.896	0.896	0.908	0.908	0.900	0.900	**0.966**
mean(ionosphere)	351	0.896	0.916	0.895	0.895	0.908	0.908	0.900	0.900	**0.966**
labor_bad	20	0.779	0.726	0.763	0.763	0.743	0.743	0.892	0.892	**1.000**
labor_good	37	0.779	0.726	0.779	0.779	0.743	0.743	0.892	0.892	**1.000**
mean(labor)	57	0.779	0.726	0.773	0.773	0.743	0.743	0.892	0.892	**1.000**
lymphography_normal	2	0.687	0.476	0.651	0.651	0.546	0.553	0.997	**1.000**	0.949
lymphography_metastases	81	0.805	0.808	0.886	0.886	0.779	0.770	0.884	0.885	**0.995**
lymphography_malign	61	0.780	0.805	0.846	0.846	0.774	0.776	0.878	0.869	**0.995**
lymphography_fibrosis	4	0.872	0.866	0.939	0.939	0.799	0.796	0.993	0.750	**1.000**
mean(lymphography)	148	0.795	0.804	0.868	0.868	0.774	0.770	0.886	0.876	**0.995**
sick_negative	3541	0.959	0.956	**0.982**	**0.982**	0.933	0.933	0.976	0.976	0.897
sick_sick	231	0.959	0.956	**0.982**	**0.982**	0.933	0.933	0.976	0.976	0.897
mean(sick)	3772	0.959	0.956	**0.982**	**0.982**	0.933	0.933	0.976	0.976	0.897
sonar_Rock	97	0.759	0.797	0.806	0.806	0.772	0.772	0.801	0.801	**0.996**
sonar_Mine	111	0.759	0.797	0.805	0.805	0.772	0.772	0.801	0.801	**0.996**
mean(sonar)	208	0.759	0.797	0.806	0.806	0.772	0.772	0.801	0.801	**0.996**
fraud_0	22109	0.989	0.986	0.981	0.981	0.985	0.985	**0.997**	**0.997**	0.988
fraud_1	11667	0.950	0.944	0.743	0.743	0.949	0.949	**0.968**	**0.968**	0.955
fraud_2	8167	0.960	0.942	0.891	0.891	0.955	0.955	**0.969**	**0.969**	0.960
fraud_3	3499	0.968	0.966	0.961	0.961	**0.977**	**0.977**	0.976	0.976	0.976
mean(fraud)	45442	0.972	0.966	0.902	0.902	0.970	0.970	**0.983**	**0.983**	0.975
kdd99_r2l	900	0.963	0.985	0.979	0.979	**0.998**	**0.998**	**0.998**	0.997	0.948
kdd99_u2r	40	0.822	0.812	0.873	0.873	**0.983**	0.976	0.976	0.976	0.923
kdd99_dos	366458	0.999	0.999	0.999	0.999	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
kdd99_probe	3707	0.992	0.990	0.994	0.994	**0.999**	**0.999**	**0.999**	**0.999**	0.949
kdd99_normal	92270	0.999	0.999	0.999	0.999	**1.000**	0.999	**1.000**	**1.000**	**1.000**
mean(kdd99)	464075	0.999	0.999	0.999	0.999	**1.000**	**1.000**	**1.000**	**1.000**	0.999

Antonie ML, Zaïane OR (2004) An Associative Classifier based on Positive and Negative Rules. In: Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2004, pp 64–69

Arunasalam B, Chawla S (2006) CCCS: A Top-Down Association Classifier for Imbalanced Class Distribution. In: Proceedings of ACM SIGKDD International Conference on Kwnoledge Discovery and Data Mining, 2006, pp 517–522

Asuncion A, Newman DJ (2007) UCI, Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, `http://www.ics.uci.edu/$\sim$mlearn/{MLR}epositoryhtml`

Bay SD, Pazzani MJ (2001) Detecting Group Differences: Mining Contrast Sets. *Data Min Knowl Disc*, 5(3): 213–246

Berger AL, Della Pietra VJ, Della Pietra SA (1996) A Maximum Entropy Approach to Natural Language Processing. *J Artif Intell Res*, 22(1): 39–71

Cesario E, Folino F, Locane A, Manco G, Ortale R (2008) Boosting Text Segmentation via Progressive Classification. Knowl Inf Syst, 15(3):285–320

Chawla NV, Bowyer KW, Hall LO and Kegelmeyer WP (2002) SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res, 16(1):321–357

Chawla NV, Lazarevic A, Hall LO and Bowyer K (2003) SMOTEBoost: Improving Prediction of Minority Class in Boosting. In: Proceedings of Principles of Knowledge Discovery in Databases, 2003, pp 107–119

Cheng H, Yan X, Han J, Hsu CW (2007) Discriminative Frequent Pattern Analysis for Effective Classification. In: Proceedings of International Conference on Data Engineering, 2007, pp 716–725

Coenen F (2004) LUCS KDD implementations of CBA and CMAR. Department of Computer Science, The University of Liverpool, UK, `http://wwwcsclivacuk/~frans/KDD/Software/`

Cohen WW (1995) Fast Effective Rule Induction. In: Proceedings of Conference on Machine Learning, 1995, pp 115–123

Cong G, Xu X, Pan F, Tung A and Yang J (2004) FARMER: Finding Interesting Rule Groups in Microarray Datasets. In: Proceedings of ACM SIGMOD International Conference on Management of data, 2004, pp 123–126

Costa G, Guarascio M, Manco G, Ortale R, Ritacco E (2009) Rule Learning with Probabilistic Smoothing. In: Proceedings of International Conference on Data Warehousing and Knowledge Discovery, 2009, pp 428–440

Duda RO, Hart PE, Stork DG (2001). Pattern Classification. Wiley & Sons. 2001

Elkan C (2001) The Foundations of Cost-Sensitive Learning. In: Proceedings of International Joint Conference on Artificial Intelligence, 2001, pp 973–978

Ezawa K, Singh M, Norton SW (1996) Learning Goal Oriented Bayesian Networks for Telecommunications Risk Management. In: Proceedings of International Conference on Machine Learning, 1996, pp 139–147

Fan W, Stolfo SJ, Zhang J and Chan PK (1999) AdaCost: Misclassification Cost-sensitive Boosting. In: Proceedings of International Conference on Machine Learning, 1999, pp 97–105

Fawcett RE, Provost F (1997) Adaptive Fraud Detection. Data Min Knowl Disc 3(1):291–316

Frank E, Witten IH (1998) Generating Accurate Rule Sets without Global Optimization. In: Proceedings of International Conference on Machine Learning, 1998, pp 144–151

Hämäläinen W (2010) StatApriori: an efficient algorithm for searching statistically significant association rules. Knowledge and Information Systems, 23(3):373-399

Han J and Yin Y (2000) Mining Frequent Patterns without Candidate Generation. In: Proceedings of ACM SIGMOD International Conference on Management of data, 2000, pp 1–12

Holte RC, Acker L, Porter B (1989) Concept Learning and the Problem of Small Disjuncts. In: Proceedings of International Conference on Artificial Intelligence, 1989, pp 813–818

Japkowicz N (2000) The Class Imbalance Problem: Significance and Strategies. In: Proceedings of International Conference on Artificial Intelligence, 2000, pp 111–117

Japkowicz N, Stephen S (2002) The class imbalance problem: A systematic study. Intell Data Anal, 6(5):429–449

Joshi MV, Agarwal RC and Kumar V (2002) Predicting Rare Classes: Can Boosting Make Any Weak Learner Strong? In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp 297–306

Joshi MV, Kumar V and Agarwal RC (2001) Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements. In: Proceedings of IEEE International Conference on Data Mining, 2001, pp 257–264

Kubat M, Holte RC, Matwin S, Kohavi R, Provost F (1998) Machine Learning for the Detection of Oil Spills in Satellite Radar Images. Mach Learn 30(2):192–215

Kubat M and Matwin S (1997) Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In: Proceedings of International Conference on Machine Learning, 1997, pp 179–186

Li W, Han J and Pei J (2001) CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules. In: Proceedings of IEEE International Conference on Data Mining, 2001, pp 369–376
Liu B, Hsu W and Ma Y (1998) Integrating Classification and Association Rule Mining. In: Proceedings of ACM SIGKDD International Conference on Kwnoledge Discovery and Data Mining, 1998, pp 80–86
Liu B, Ma Y, Wong CK (2000) Improving an Association Rule Based Classifier. In: Proceedings of Principles of Data Mining and Knowledge Discovery, 2000, pp 504–509
McLachlan G and Peel D (2000). Finite Mixture Models. Wiley, 2000
Mitchell TM (1997). Machine Learning. McGraw-Hill, 1997
Pazzani M, Merz C, Murphy P, Hume T, Brunk C (1994) Reducing Misclassification Costs. In: Proceedings of International Conference on Machine Learning, 1994, pp 217–225
Phua C, Alahakoon D, Lee V (2004) Minority Report in Fraud Detection: Classification of Skewed Data. ACM SIGKDD Explorations Newsletter. Special issue on learning from imbalanced datasets:50–59
Provost F and Fawcett T (2001) Robust Classification for Imprecise Environments. Mach Learn. 42(3):203–231
Quinlan JR and Cameron-Jones RM (1993) FOIL: A Midterm Report. In: Proceedings of European Conference on Machine Learning, 1993, pp 3–20
Riddle P, Segal R, Etzioni O (1994) Representation Design and Brute-force Induction in a Boeing Manufacturing Domain. Appl Artif Intell, 8(1):125–147
Tang J, Chen Z, Fu A and Cheung D (2007) Capabilities of outlier detection schemes in large datasets, framework and methodologies. Knowl Inf Sys, 11(1):45-84
Tatti N (2008) Maximum entropy based significance of itemsets. Knowl Inf Sys, 17(1):57-77
Thabtah F (2007) A Review of Associative Classification Mining. J Knowl Eng Rev, 22(1):37–65
Ting KM (2000) A Comparative Study of Cost-Sensitive Boosting Algorithms. In: Proceedings of International Conference on Machine Learning, 2000, pp 983–990
Wang J and Karypis G (2005) HARMONY: Efficiently Mining the Best Rules for Classification. In: Proceedings of SIAM International Conference on Data Mining, 2005, pp 205–216
Webb G, Boughton J and Wang Z (2005) Not so naive Bayes: Aggregating one-dependence estimators. Mach Learn, 58(1): 5–24
Weiss GM (2000) Learning with Rare Cases and Small Disjuncts. In: Proceedings of International Conference on Machine Learning, 2000, pp 558–565
Weiss GM (2004) Mining with Rarity: A Unifying Framework. ACM SIGKDD Explorations Newsletter, 6(1):7–19
Weiss GM, Hirsh H (2000) A Quantitative Study of Small Disjuncts. In: Proceedings of National Conference on Artificial Intelligence, 2000, pp 665–670
Weiss GM and Provost F (2003) Learning when Training Data are Costly: the Effect of Class Distribution on Tree Induction. J Artif Intell Res, 19:315–354
Xin X and Han J (2003) CPAR: Classification based on Predictive Association Rules. In: Proceedings of SIAM International Conference on Data Mining, 2003, pp 331–335

Author Biographies

Gianni Costa is currently researcher at the Institute of High Performance Computing and Networks (ICAR-CNR) of the National Research Council of Italy. He graduated summa cum laude in computer science engineering in 2003 and recieved Ph.D. in systems and computer engineering in 2007 from the University of Calabria, Italy. From 2003 to 2006 he was a Ph.D. student at University of Calabria. His current research interests include the following fields: data mining and knowledge discovery, semistructured data, entity resolution.

Giuseppe Manco graduated summa cum laude in computer science in 1994 and received the PhD degree in computer science from the University of Pisa. He is currently a senior researcher at the Institute of High Performance Computing and Networks (ICAR-CNR) of the National Research Council of Italy and a con- tract professor at University of Calabria, Italy. He has been contract researcher at the CNUCE Institute in Pisa, Italy, and a visiting fellow at the CWI Institute in Amsterdam, Nederlands. His current research interests include deductive databases, knowledge discovery and data mining, Web databases, and semistructured data.

Riccardo Ortale is researcher at the Institute of High Performance Computing and Networks (ICAR-CNR) of the National Research Council of Italy.is researcher at the Institute of High Performance Computing and Networks (ICAR-CNR) of the National Research Council of Italy is researcher at the Institute of High Performance Computing and Networks (ICAR-CNR) of the National Research Council of Italy is researcher at the Institute of High Performance Computing and Networks (ICAR-CNR) of the National Research Council of Italy ....

Ettore Ritacco received the M.S. and Ph.D degree in computer science in 2006 and 2010, respectively, from the University of Calabria (UNICAL). Currently he is a research fellow at the Institute of High Performance Computing and Networks (ICAR-CNR) of the National Research Council of Italy. His research interests include data mining and machine learning.

Correspondence and offprint requests to: Riccardo Ortale, ICAR-CNR, Via P. Bucci 41c - 87036 Rende (CS), Italy. Email: ortale@icar.cnr.it