Competence-Conscious Associative Classification

Adriano Veloso[a], Mohammed Zaki[b], Wagner Meira Jr.[a] and Marcos Gonçalves[a]
[a] Computer Science Dept, Federal University of Minas Gerais, Brazil
{adrianov,meira,mgoncalv}@dcc.ufmg.br
[b] Computer Science Dept, Rensselaer Polytechnic Institute, Troy, USA
zaki@cs.rpi.edu

Abstract

The classification performance of an associative classifier is strongly dependent on the statistic measure or metric that is used to quantify the strength of the association between features and classes (i.e., confidence, correlation etc.). Previous studies have shown that classifiers produced by different metrics may provide conflicting predictions, and that the best metric to use is data-dependent and rarely known while designing the classifier. This uncertainty concerning the optimal match between metrics and problems is a dilemma, and prevents associative classifiers to achieve their maximal performance. This dilemma is the focus of this paper.

A possible solution to this dilemma is to learn the competence, expertise, or assertiveness of metrics. The basic idea is that each metric has a specific sub-domain for which it is most competent (i.e., it consistently produces more accurate classifiers than the ones produced by other metrics). Particularly, we investigate stacking-based meta-learning methods, which use the training data to find the domain of competence of each metric. The meta-classifier describes the domains of competence (or areas of expertise) of each metric, enabling a more sensible use of these metrics so that competence-conscious classifiers can be produced (i.e., a metric is only used to produce classifiers for test instances that belong to its domain of competence). We conducted a systematic and comprehensive evaluation, using different datasets and evaluation measures, of classifiers produced by different metrics. The result is that, while no metric is always superior than all others, the selection of appropriate metrics according to their competence/expertise (i.e., competence-conscious associative classifiers) seems very effective, showing gains that range from 1.2% to 26.3% when compared to the baselines (SVMs and an existing ensemble method).

1 Introduction

One strategy for devising a classifier is to exploit relationships, dependencies and associations between features and classes. Such associations are usually hidden in the training examples, and when uncovered, they may reveal important aspects concerning the underlying phenomenon that generated these examples. These aspects can be expressed using rules of the form $\mathcal{X} \rightarrow c$, which indicate that a set of features $\mathcal{X}$ is associated with class c. The use of such rules for the sake of prediction has led to a new family of classifiers, that are often referred to as associative classifiers [4, 8–10, 21, 22, 26, 27, 29, 33, 37]. These classifiers have shown to be valuable in many applications, including gene functional analysis [35], document categorization [34, 39], Web ranking [36] etc.

An obvious difference between various associative classifiers resides in the metric used to capture dependencies between features and classes (i.e., confidence, correlation etc.). It has been observed that many such metrics provide conflicting information about feature-class dependencies, resulting in associative classifiers with different classification performances [34]. In fact, different metrics have different intrinsic properties [28] (i.e., symmetric/asymmetric metrics, scaling variant/invariant metrics etc.), and this may lead to very different associative classifiers. Some of these classifiers may be well suited for some classification problems, but not for others (that is, each metric has a particular domain for which it is more competent). Competent metrics are rarely known while devising the classifier, and this dilemma concerning the best match between metrics and problems prevents the full potential of associative classifiers.

Obviously, one possible solution to the metric dilemma is to find the domain of competence (or areas of expertise) for each metric, that is, subsets of examples for which a certain metric produces better classifiers than the others. Having this information would enable the assignment of competent associative classifiers to specific problems according to their competence/expertise [12, 23]. Hopefully, classification performance would be drastically boosted by taking advantage of consciously assigning metrics to specific subsets of instances (i.e., a domain of competence). This would be great, except that there are several metrics, and numerous (unknown) characteristics affecting their corresponding competence, and finding such an invariant domain of com-

petence for metrics seems to be practically unfeasible.

As an alternate approach to the metric dilemma, we propose to automatically extract the competence of each metric. Taking as a starting point a set of q *accurate*[1] and *diverse*[2] metrics, m_1, m_2, ..., m_q, a stacking-like meta-learning strategy [30, 38], which is a procedure based on the idea that different classifiers may provide different but complementary explanations of the data, is used to extract, from the training data, information regarding the competence of each metric. This information is then used to produce the meta-data. Specifically, it is explicitly indicated the metrics that correctly classify each example in the training data (i.e., using a cross-validation procedure).

The meta-data is used to produce a meta-classifier which has the ability to consciously decide the appropriate match between metrics and examples (i.e., the meta-classifier is a function mapping features to competent metrics). Then, for each test instance t, the meta-classifier is used to decide which is the most competent metric to be applied, according to their expertise. A specific classifier, $\mathcal{C}^t_{m_i}$, is finally produced, so that m_i is expected to be the most competent metric to classify instance t (i.e., t belongs to the domain of competence of metric m_i). The classifiers that are produced following this strategy are regarded as *competence-conscious* classifiers. We propose two competence-conscious classifiers, with the difference between them residing in the way they perform the analysis of the domains of competence (or areas of expertise). The first classifier performs a class-centric analysis, in which the domain of competence of a metric is composed of *classes* for which it produces accurate classifiers. The other classifier performs a different analysis, in which the domain of competence of a metric is composed of *examples* for which it produces accurate classifiers.

To evaluate the effectiveness of competence-conscious associative classifiers, we performed a systematic set of experiments using the UCI datasets, as well as more complex datasets obtained from other real applications, such as digital libraries, Web directories and Web spam detection. Our results suggest that the finer-grained the analysis of the domains of competence (i.e., from classes to instances), the more effective is the final associative classifier. The results also show that the proposed competence-conscious associative classifiers are able to outperform the baselines (SVMs and existing ensemble methods), providing gains ranging from 1.2% to 26.3%. In sum, the specific contributions of this paper are:

- We present a comprehensive study of the competence of associative classifiers produced by different statistic metrics. We show that no metric is consistently better than the others for all problems. Further, we also show that traditional metrics, such as confidence, are just moderately competent for most of the problems investigated.

- We propose competence-conscious classifiers, which effectively combine classifiers produced by different metrics (an ensemble) using their domains of competence. All constituent classifiers are produced using the same rule set. The only difference between the base classifiers is the way they interpret the rules (each classifier employes a different metric).

- We used several complex datasets to present a deep evaluation of the proposed competence-conscious classifiers, and we show that they are able to provide expressive gains in classification performance. Our analysis include a study about how the diversity and the accuracy of the base classifiers affect the performance of competence-conscious associative classifiers.

The remaining of this paper is organized as follows. In Section 2 we discuss related work. Then, in Section 3, we introduce our associative classification technique. The metric dilemma, and competence-conscious associative classifiers (i.e., the ensemble of classifiers produced by different metrics), are presented in Section 4. In Section 5 we evaluate the proposed competence-conscious classifiers, and compare them against state-of-the-art SVMs and existing ensemble techniques. Finally, in Section 6 we conclude the paper.

2 Related Work

The ultimate goal of a classifier is to achieve the best possible classification performance for the problem at hand. An ensemble is a collection of classifiers whose predictions are combined with the goal of achieving better performance than the constituent classifiers. There is a body of evidence suggesting that ensembles offer substantial advantages in enough situations to be regarded as a major advance in machine learning [14]. Also, there is a body of theory explaining why ensembles work. In this section we will discuss approaches used to produce ensembles, and how they relate to the approaches presented in this paper.

A variety of ensemble methods has already been proposed. Well known methods include bagging [6], boosting [24], and stacking [38]. Bagging is the acronym for bootstrap aggregating. It generates bootstrap replicates of the training data by sampling examples from the original training data uniformly and with replacement (that is, an example may appear repeated times or not at all in any particular replicate). A classification model is produced from each replicate, and then they are combined using approaches

[1] An accurate metric is one that produces a classifier that has an error rate of better than random guessing

[2] Two metrics are diverse if they produce classifiers that misclassify different instances

such as averaging or voting [19]. Boosting combines several weak classifiers (those that are only slightly correlated with the true classification) in order to produce a single and much stronger one. Combination approaches include averaging [18] and majority [13]. In the following, we will focus our attention on stacking methods, since the techniques proposed in this paper are mostly related to them. Stacking is based on the idea that different classifiers may provide different but complementary explanations of the data. Thus, the predictions of these different (base) classifiers can potentially produce novel information that can be used as meta-features to form a new training data. Then, a meta-classifier is built using this new training data, but instead of predicting the correct class for a given test instance, the meta-classifier predictcs the base classifier that is most likely to correctly predict the class for such instance. The obvious advantage, in this case, is that the errors of a base classifier can be counterbalanced by the hits of others.

In this paper we are interested in associative classification. We exploit stacking based meta-learning approaches to address an important issue in associative classification: the *metric dilemma*. Several statistic metrics can be used to estimate *feature-class* associations [16, 20, 28], but the most competent one is rarely known in advance. Thus, we propose to explore the diversity among classifiers that are produced using different statistic metrics to maximize the performance of the final classifier (which will be refereed as a competence-conscious associative classifier). The metric dilemma is challenging, and, as far as we know, this is the first attempt to integrate classifiers produced by different statistic metrics, in the context of associative classification.

The integration of classifiers using strategies related to stacking was largely explored [2, 12, 15, 23, 32]. In [2], the authors use a neural network to learn, from predefined meta-features (e.g., maximum confidence, average confidence, number of applicable rules etc.), how to weigh the rules using a single association metric (i.e., confidence). We believe that the work of Ortega et. al [23] is the closest to ours. They used a referee (which in our case is a meta-classifier) to indicate the best classifier to be applied for each example. *In our experiments we performed a direct comparison between the competence-conscious classifiers proposed in this paper and the ensemble approach proposed in* [23].

Self-delegation [12] is another strategy for combining the predictions of different base classifiers, and thus it is also related to our work. The idea is that each base classifier chooses by itself which instances it can safely classify. This choice is based on the confidence in its prediction. A base classifier delegates the difficult or uncertain predictions to other classifiers. Clearly, this strategy produces classifiers which are exclusively defined in terms of the original features (no meta-features are generated). This simplicity may be desirable, but it may neglect important information associated with meta-features. *We show this by performing a direct comparison between self-delegating classifiers and competence-conscious classifiers*. We concluded that sensitive information regarding the competence of metrics leads to associative classifiers that provide substantial improvements in classification performance.

3 Associative Classification

The classification problem is defined as follows. We have an input dataset called the *training data* (denoted as $\mathcal{D}$) which consists of instances composed of a set of l attribute-values $(a_1, a_2, \ldots, a_l)$ along with a special variable called the *class*. The set of all possible attribute-values is denoted as $\mathcal{A}$, while the class variable draws its value from a discrete set of classes $(c_1, c_2, \ldots, c_n)$. The training data is used to build a classifier that relates features (or attribute values) to the class variable. The *test instances* are a set of instances for which only the features are known while the class value is unknown. The classifier, which is a function from $\mathcal{A}$ to $\{c_1, c_2, \ldots, c_n\}$, is used to predict the class value for test instances (i.e., the classifier is a function which maps a set of features to one of the classes).

Associative classifiers exploit the fact that, frequently, there are strong associations between features and classes. Typically, such associations are expressed using rules of the form $\mathcal{X} \rightarrow c_i$, where $\mathcal{X} \subseteq \mathcal{A}$ and c_i is one of the classes. These rules are usually hidden in the training data, and when uncovered they can be combined in order to accurately map features to classes (i.e., the classification function is obtained by combining the information provided by these rules). In the following we will denote as $\mathcal{R}$ an arbitrary rule set extracted from $\mathcal{D}$. Similarly, we will denote as $\mathcal{R}_{c_i}$ an arbitrary rule set composed of rules of the form $\mathcal{X} \rightarrow c_i$, such that $\mathcal{R}_{c_i} \subseteq \mathcal{R}$.

Naturally, some rules in $\mathcal{R}$ represent stronger associations than others. A set of statistic metrics that quantify the strength of the association between $\mathcal{X}$ and c_i are used to compare the rules. Associative classifiers usually learn the classification function in two broad steps:

1. generate a rule set, $\mathcal{R}$, from $\mathcal{D}$
2. estimate the likelihood of class membership for each test instance, by combining the information provided by rules in $\mathcal{R}$

The main challenge associated with the first step is to avoid the rule explosion problem, which may make the rule extraction process unfeasible in practice (or at least time-consuming). An approach to deal with this challenge (the pruning dilemma) was proposed in [33]. In this paper we follow this approach [33], which generates rules on a demand-driven basis according to each test instance. This

strategy drastically reduces the number of extracted rules, while at the same time, it reduces the chance of missing important rules.

In this paper our focus is on an important challenge associated with step 2: provide an accurate estimate of the likelihood of class membership. Specifically, given an instance t, we want to estimate the likelihood $\hat{p}(c_i|t)$ that t belongs to class c_i. Only rules $\mathcal{X} \rightarrow \{c_1, c_2 \ldots c_n\} \in \mathcal{R}$, such that $\mathcal{X} \subseteq t$ are used to estimate $\hat{p}(c_1|t), \ldots \hat{p}(c_n|t)$. Such rules are said to match t, and they form the rule set $\mathcal{R}^t$. Obviously, $\mathcal{R}^t \subseteq \mathcal{R}$.

The likelihood of membership of an instance t is estimated by combining rules in $\mathcal{R}^t$=$\{\mathcal{R}^t_{c_1} \cup \mathcal{R}^t_{c_2} \cup \ldots \cup \mathcal{R}^t_{c_n}\}$. A simple (yet effective) probabilistic strategy is to interpret $\mathcal{R}^t$ as a poll, in which rule $\mathcal{X} \rightarrow c_i \in \mathcal{R}^t_{c_i}$ is a vote given by $\mathcal{X}$ for class c_i. The weight of a vote $\mathcal{X} \xrightarrow{m} c_i$ depends on the strength of the association between $\mathcal{X}$ and c_i, which is given by an association metric m. Weighted votes for class c_i are summed and then averaged by the total number of votes for this class, as expressed by function $s(c_i, t)$, shown in Equation 3.1 (where $m(r)$ is the metric value for rule r). As will be discussed in the next section, there are cases in which $m(r) < 0$, and thus a value z (which is the lowest score, that is, z=$s(c_j, t)|s(c_j, t) \leq s(c_i, t) \forall c_i$), is used to ensure that all scores are greater than or equal to 0.

$$s(c_i, t) = \frac{\sum_{r \in \mathcal{R}^t_{c_i}} m(r)}{\mid \mathcal{R}^t_{c_i} \mid} - z \tag{3.1}$$

The likelihood of membership of t to class c_i is expressed by the function $\hat{p}(c_i|t)$, shown in Equation 3.2 (thus, votes with high weights increase the likelihood of the corresponding class being the correct one, while votes with low weights reduce the likelihood of the corresponding class being the correct one). A higher value of $\hat{p}(c_i|t)$ indicates a higher likelihood of t to belong to class c_i. The class associated with the highest likelihood is finally predicted. As will be shown in Section 5, association metrics play a fundamental role in estimating the likelihood of class membership. However, the best-quality, most competent metric is data-dependent, and rarely known while devising the classifier[3].

$$\hat{p}(c_i|t) = \frac{s(c_i, t)}{\sum_{j}^{n} s(c_j, t)} \tag{3.2}$$

[3]We denote as $\mathcal{C}_{m_j}$ an associative classifier which applies m_j as the association metric in Equation 3.1

4 The Metric Dilemma

Selecting an appropriate association metric is a major issue while designing an associative classifier. Classifiers produced by different metrics often present different classification performance. Depending on the characteristics of the problem, some metrics may be more suitable than others. That is, a sub-domain may present properties that make a metric more suitable than others. This suggests that classifiers produced by a certain metric are only able to make reliable predictions over a subset of the entire domain space, which is the area of expertise, or domain of competence, of such metric. In this section we exploit the training data to learn the competence, or expertise of each metric. Then, a specific metric is used to produce a classifier for subproblems that belong to its domain of competence.

4.1 Association Metrics Next we present several metrics for measuring the strength of association between a set of features ($\mathcal{X}$) and classes ($c_1, c_2, \ldots, c_n$). Some of these metrics are popular ones in routine use [1,28], while others were recently used in the context of associative classification [3]. Naturally, there is nothing to gain by combining similar classifiers (i.e., those that perform almost the same predictions), and thus we selected metrics that are sufficiently different (these metrics have different properties according to [28]) to indicate that the corresponding classifiers ($\mathcal{C}_{m_1}, \ldots, \mathcal{C}_{m_8}$) may present some diversity.

- Confidence (m_1) [1]: This metric measures the fraction of instances in $\mathcal{D}$ containing $\mathcal{X}$ that belong to c_i. It is the conditional probability of c_i being the correct class of instance t given that $\mathcal{X} \subseteq t$, as shown in Equation 4.3. Its value ranges from 0 to 1.

$$m_1 = p(c_i|\mathcal{X}) \tag{4.3}$$

- Added Value (m_2) [16]: This metric measures the gain in accuracy obtained by using rule $\mathcal{X} \rightarrow c_i$ instead of always predicting c_i, as shown in Equation 4.4. Negative values indicate that always predicting c_i is better than using the rule. Its value ranges from -1 to 1.

$$m_2 = p(c_i|\mathcal{X}) - p(c_i) \tag{4.4}$$

- Certainty (m_3) [20]: This metric measures the increase in accuracy between rule $\mathcal{X} \rightarrow c_i$ and always predicting c_i, as shown in Equation 4.5. It assumes values smaller than 1.

$$m_3 = \frac{p(c_i|\mathcal{X}) - p(c_i)}{p(\overline{c_i})} \tag{4.5}$$

- Yules'Q (m_4) and Yules'Y (m_5) [40]: These metrics are based on odds value, as shown in Equations 4.6 and 4.7, respectively. Their values range from -1 to 1. The value 1 implies perfect positive association between $\mathcal{X}$ and c_i, value 0 implies no association, and value -1 implies perfect negative association.

(4.6)
$$m_4 = \frac{p(\mathcal{X} \cup c_i)p(\overline{\mathcal{X} \cup c_i}) - p(\mathcal{X} \cup \overline{c_i})p(\overline{\mathcal{X}} \cup c_i)}{p(\mathcal{X} \cup c_i)p(\overline{\mathcal{X} \cup c_i}) + p(\mathcal{X} \cup \overline{c_i})p(\overline{\mathcal{X}} \cup c_i)}$$

(4.7)
$$m_5 = \frac{\sqrt{p(\mathcal{X} \cup c_i)p(\overline{\mathcal{X} \cup c_i})} - \sqrt{p(\mathcal{X} \cup \overline{c_i})p(\overline{\mathcal{X}} \cup c_i)}}{\sqrt{p(\mathcal{X} \cup c_i)p(\overline{\mathcal{X} \cup c_i})} + \sqrt{p(\mathcal{X} \cup \overline{c_i})p(\overline{\mathcal{X}} \cup c_i)}}$$

- Strength Score (m_6) [3]: This metric measures the correlation between $\mathcal{X}$ and c_i, but it also takes into account how $\mathcal{X}$ is correlated with the complement of c_i (i.e., $\overline{c_i}$), as shown in Equation 4.8. Its value ranges from 0 to ∞.

(4.8)
$$m_6 = \frac{p(\mathcal{X}|c_i)p(c_i|\mathcal{X})}{p(\mathcal{X}|\overline{c_i})}$$

- Support (m_7) [1]: This metric measures the fraction of instances in $\mathcal{D}$ covered by the rule $\mathcal{X} \rightarrow c_i$, as shown in Equation 4.9. Its value ranges from 0 to 1.

(4.9)
$$m_7 = p(\mathcal{X} \cup c_i)$$

- Weighted Relative Confidence (m_8) [20]: This metric trades off accuracy and generality, as shown in Equation 4.10. The first component is the accuracy gain that is obtained by using rule $\mathcal{X} \rightarrow c_i$ instead of always predicting c_i. The second component incorporates generality.

(4.10)
$$m_8 = (p(c_i|\mathcal{X}) - p(c_i))p(\mathcal{X})$$

Although we focus our analysis only on these eight metrics, the approaches to be introduced here are general and able to exploit any number of metrics transparently.

Associative classifiers produced by different metrics may perform different predictions. Table 1 shows a simple example containing a training data from which different associative classifiers may be produced in order to predict the class for instance 11. For simplicity, we will consider only attributes a_1, a_2 and a_l. In this case, according to Equation 3.2, classifier $\mathcal{C}^{11}_{m_1}$ predicts c_3, classifiers $\mathcal{C}^{11}_{m_2}$, $\mathcal{C}^{11}_{m_3}$, and $\mathcal{C}^{11}_{m_8}$ predict c_1, and classifiers $\mathcal{C}^{11}_{m_4}$, $\mathcal{C}^{11}_{m_5}$, $\mathcal{C}^{11}_{m_6}$, and $\mathcal{C}^{11}_{m_7}$ predict c_2. Next we will discuss a simple approach to boost classification performance by exploiting associative classifiers produced by the aforementioned metrics.

Id	Class	Attribute-Values $a_1\ a_2 \ldots a_l$
1	c_1	1 3 ... 6
2	c_1	1 3 ... 7
3	c_1	2 4 ... 6
4	c_2	2 4 ... 7
5	c_2	2 5 ... 8
6	c_2	2 4 ... 6
7	c_3	1 3 ... 9
8	c_3	2 5 ... 9
9	c_3	2 4 ... 8
10	c_3	2 4 ... 9
11	c_2	2 3 ... 8

Table 1: Training data, $\mathcal{D}$ (first 10 instances), and test set, $\mathcal{T}$ (instance 11).

Self-Delegating Classifier (SDC) Equation 3.2 can be used to estimate the reliability of a prediction, and this information can be used to select the most reliable prediction from all involved classifiers [12]. The process is illustrated in Algorithm 1. For a given test instance t, the selected class is the one which is associated with the highest likelihood $\hat{p}(c_i|t)$ amongst all classifiers $\mathcal{C}^t_{m_1}, \mathcal{C}^t_{m_2}, \ldots, \mathcal{C}^t_{m_q}$. The basic idea is to use the most reliable prediction (among the predictions performed by all classifiers) to select the class for instance t.

Although simple, SDC does not exploit the competence of each metric. In fact, each base classifier simply decides by itself the instances it will classify, not meaning that the select instances belong to its domain of competence.

Algorithm 1 Classifier based on self delegation of metrics.

Require: The training data $\mathcal{D}$, and a test instance t
Ensure: The class for instance t

1: $\mathcal{R}^t \Leftarrow$ rules $\mathcal{X} \rightarrow c_i$ (with $1 \leq i \leq n$) extracted from $\mathcal{D}$ such that $\mathcal{X} \subseteq t$
2: produce different classifiers $\mathcal{C}^t_{m_1}, \mathcal{C}^t_{m_2}, \ldots, \mathcal{C}^t_{m_q}$, for instance t, using rules in $\mathcal{R}^t$
3: **return** the class associated with the highest likelihood of membership for t (i.e., Eq. 3.2), amongst all classifiers

4.2 Learning the Metric Competence The optimal match between metrics and problems is valuable information. In this section we present an approach to estimate such matching. The proposed approach may be viewed as an application of Wolpert's stacked generalization [38]. From a general point of view, stacking can be considered a meta-learning method, as it refers to the induction of classifiers over inputs that are, in turn, the predictions of other classifiers induced from the training data (i.e., meta-data). Specif-

Id	Class	Attribute-Values $a_1\ a_2\ \ldots\ a_l$	Competent Metric(s) (per instance)	Most Competent Metric(s) (per class)
1	c_1	1 3 ... 6	m_2	m_1
2	c_1	1 3 ... 7	m_1 m_3	
3	c_1	2 4 ... 6	m_1	
4	c_2	2 4 ... 7	m_1 m_2	m_1
5	c_2	2 5 ... 8	m_1 m_2 m_3	
6	c_2	2 4 ... 6	m_1	
7	c_3	1 3 ... 9	m_2	m_2
8	c_3	2 5 ... 9	m_2 m_3	
9	c_3	2 4 ... 8	m_1 m_2 m_3	
10	c_3	2 4 ... 9	m_2	

Table 2: Enhanced training data, $\mathcal{D}_e$.

ically, in the stacking strategy an algorithm is applied on meta-data, which can contain properties of the original data, properties of different classification algorithms, or patterns previously derived from the original data. The meta-data is used to select, alter, or combine different classifiers, which, in our case, are classifiers produced by different metrics

Algorithm 2 Enhancing the training data with the competence of each metric.

Require: The original training data $\mathcal{D}$, and a cross-validation parameter k

Ensure: The enhanced training data $\mathcal{D}_e$

1: split $\mathcal{D}$ into k partitions, so that $\mathcal{D}$=$\{d_1 \cup d_2 \cup \ldots \cup d_k\}$
2: $\mathcal{D}_e \Leftarrow \emptyset$
3: **for each** partition d_i **do**
4: **for each** instance $t \in d_i$ **do**
5: $m \Leftarrow \emptyset$
6: $\mathcal{R}^t \Leftarrow$ rules $\mathcal{X} \rightarrow c_y$ (with $1 \leq y \leq n$) extracted from $\{\mathcal{D}\text{-}d_i\}$ such that $\mathcal{X} \subseteq t$
7: produce different classifiers, $\mathcal{C}^t_{m_1}, \mathcal{C}^t_{m_2}, \ldots, \mathcal{C}^t_{m_q}$, using rules in $\mathcal{R}^t$
8: **for each** classifier $\mathcal{C}^t_{m_j}$ **do**
9: **if** $\mathcal{C}^t_{m_j}$ correctly predicts the class for t **then**
10: $m \Leftarrow m \cup m_j$
11: **end if**
12: **end for**
13: $\mathcal{D}_e \Leftarrow \mathcal{D}_e \cup \{t \cup m\}$
14: **end for**
15: **end for**

The process starts by enhancing the original training data (i.e., producing the meta-data) using the outputs of the base classifiers, $\mathcal{C}^t_{m_1}$, $\mathcal{C}^t_{m_2} \ldots \mathcal{C}^t_{m_q}$. Algorithm 2 shows the basic steps involved in the process. Initially, the enhanced training data, $\mathcal{D}_e$ is empty. An example t, along with the competence of each metric with regard to t (i.e., which metric correctly predicted the class for t), is inserted into $\mathcal{D}_e$. The process continues until all examples are processed. In the end, for each example $t \in \mathcal{D}_e$ we have a list of metrics that produced a competent classifier for t, and this information enables learning the domains of competence of each metric, as will be discussed in the next section.

To illustrate this process, please consider the example shown in Tables 1 and 2. Table 1 shows the original training data, $\mathcal{D}$. Using the process described in Algorithm 2, the competence of each metric to each instance is appended to $\mathcal{D}$, resulting in the enhanced training data, $\mathcal{D}_e$, which is shown in Table 2. In this case, for a given example t, metric m_i is shown if the corresponding classifier $\mathcal{C}^t_{m_i}$ has correctly classified t using the stacking procedure (i.e., metric m_i is competent with regard to example t). The enhanced training data, $\mathcal{D}_e$, can be exploited in several ways. In particular, we will use $\mathcal{D}_e$ to produce competence-conscious classifiers, as will be discussed next.

4.3 Competence-Conscious Classifiers In this section we present strategies for exploiting $\mathcal{D}_e$ in order to produce competence-conscious associative classifiers. The challenge, in this case, is to properly select a competent metric for a specific test instance. The competence-conscious classifiers to be presented differ in how they perform the analysis of the domains of competence of metrics.

Class-Centric Competence-Conscious Classifier (C^5) The competence of a metric is often associated with certain classes. Some metrics, for instance, produce classifiers which show preference for more frequent classes, while others produce classifiers which show preference for less frequent ones. As an illustrative example, please consider Table 2. Metric m_1 is extremely competent for classifying instances that belong to classes c_1 and c_2. On the other hand, if we consider instances belonging to c_3, metric m_2 perfectly classifies all instances. This information (which is shown in the last column of Table 2) may be used to produce class-centric competence-conscious classifiers. The process

is depicted in Algorithm 3. It starts with a meta-classifier, $\mathcal{M}$, which learns the most competent metric for a given class. Any classifier can be used to build the meta-classifier. For simplicity we choose an associative classifier that weights the votes given by rules using the confidence metric. In this case, instead of generating rules $\mathcal{X} \rightarrow c_i$, the meta-classifier generates rules $\mathcal{X} \rightarrow m_i$, which maps features (i.e., in the third column of Table 2) to metrics (i.e., in the fifth column of Table 2). Then, for each test instance t, the meta-classifier indicates the most competent metric, m_j, that is then used to produce the final classifier, $\mathcal{C}^t_{m_j}$, which is finally used to predict the class for instance t.

Algorithm 3 Class-centric meta classifier.

Require: The enhanced training data $\mathcal{D}_e$ (i.e., the 3^{rd} and 5^{th} columns of Table 2), and a test instance t

Ensure: The most competent metric for instance t

1: **for each** metric m_i **do**
2: $\mathcal{R}^t_{m_i} \Leftarrow$ rules $\mathcal{X} \rightarrow m_i$ extracted from $\mathcal{D}_e$ such that $\mathcal{X} \subseteq t$
3: Estimate $\hat{p}(m_i|t)$, according to Equation 3.2 (using confidence to weigh the votes)
4: **end for**
5: **return** metric m_j such that $\hat{p}(m_j|t) > \hat{p}(m_i|t) \forall i \neq j$

Instance-Centric Competence-Conscious Classifier (IC^4) Although the competence of some metrics are associated with certain classes, specific instances may be better classified using other metrics. In such cases, a finer-grained analysis of competence is desired. As an illustrative example, please consider again Table 2. Although metric m_1 is the most competent one to classify instances belonging to class c_1, metric m_2 is the only one which competently classifies instance 1 (which belongs to c_1). Again, a meta-classifier, $\mathcal{M}$, is used to explore such cases. The process is depicted in algorithm 4. In this case, the meta-classifier learns the most competent metric by generating rules of the form $\mathcal{X} \rightarrow m_i$, which maps features (i.e., the third column of Table 2) to metrics (i.e., in the fourth column of Table 2). Then, for each test instance t, the meta-classifier indicates the most competent metric, m_j, which is used to produce the final classifier, $\mathcal{C}^t_{m_j}$.

The main advantage of C^5 and IC^4 is that, in practice, multiple metrics produce competent classifiers for a particular instance t, but $\mathcal{M}$ needs to predict only one of them (competent metrics are not mutually exclusive, and thus, in practice, multiple metrics produce competent classifiers for instance t). We will show in Section 5 that this redundancy in competence that exists when different metrics are taken into account, may increase the chance of selecting a competent metric.

Algorithm 4 Instance-centric meta classifier.

Require: The enhanced training data $\mathcal{D}_e$ (i.e., the 3^{rd} and 4^{th} columns of Table 2), and a test instance t

Ensure: The most competent metric for instance t

1: **for each** metric m_i **do**
2: $\mathcal{R}^t_{m_i} \Leftarrow$ rules $\mathcal{X} \rightarrow m_i$ extracted from $\mathcal{D}_e$ such that $\mathcal{X} \subseteq t$
3: Estimate $\hat{p}(m_i|t)$, according to Equation 3.2 (using confidence to weigh the votes)
4: **end for**
5: **return** metric m_j such that $\hat{p}(m_j|t) > \hat{p}(m_i|t) \forall i \neq j$

Algorithm 5 summarizes the basic steps of the competence-conscious associative classifiers. The algorithm receives a test-instance t and a meta-classifier (either C^5 or IC^4) as input, and returns as output the class associated with t.

Algorithm 5 Competence-conscious classifiers.

Require: The training data $\mathcal{D}$, the meta-classifier $\mathcal{M}$, and a test instance t

Ensure: The class for instance t

1: **for each** class c_i **do**
2: $\mathcal{R}^t_{c_i} \Leftarrow$ rules $\mathcal{X} \rightarrow c_i$ extracted from $\mathcal{D}$ such that $\mathcal{X} \subseteq t$
3: **end for**
4: select the most competent classifier for t, $\mathcal{C}^t_{m_x}$, using $\mathcal{M}$
5: Estimate $\hat{p}(c_i|t)$ (with $1 \leq i \leq n$), according to Equation 3.2 (using metric m_x to weigh the votes)
6: **return** class c_j such that $\hat{p}(c_j|t) > \hat{p}(c_i|t) \forall i \neq j$

Bounds for Competence-Conscious Classifiers We derived lower and upper bounds for the classification performance of the proposed competence-conscious associative classifiers. The lower bound is the performance that is obtained by randomly selecting a competent metric. Clearly, this lower bound increases with the redundancy between the base classifiers, $\mathcal{C}^t_{m_1}, \ldots \mathcal{C}^t_{m_q}$. The upper bound is the classification performance that would be obtained by an oracle which always predicts a competent metric (note that perfect performance is not always possible, since it may not exist a competent metric for some instances). Clearly, this upper bound increases with the accuracy and diversity associated with base classifiers.

Diversity between Base Classifiers We use the plain disagreement [31] as a measure of diversity between two base associative classifiers, $\mathcal{C}_{m_i}$ and $\mathcal{C}_{m_j}$, as shown in Equation 4.11, where $\mathcal{C}_{m_i}(t)$ is the class predicted by classifier $\mathcal{C}_{m_i}$ for instance t, and diff$(\mathcal{C}_{m_i}(t), \mathcal{C}_{m_j}(t))$ returns 1 if the two arguments are different (and 0 otherwise). The diversity

of the ensemble is finally given by averaging $\text{PD}(\mathcal{C}_{m_i}, \mathcal{C}_{m_j})$ over all q base classifiers.

$$\text{PD}(\mathcal{C}_{m_i}, \mathcal{C}_{m_j}) = \frac{1}{|\mathcal{T}|} \sum_{t=1}^{|\mathcal{T}|} \text{diff}(\mathcal{C}_{m_i}(t), \mathcal{C}_{m_j}(t)) \quad (4.11)$$

5 Experimental Evaluation

In this section we will empirically analyze the proposed classifiers, SDC, C^5, and IC^4. In our experiments, we used 26 datasets from the UCI Machine Learning Repository [5] and three datasets obtained from more complex applications. These datasets cover a wide range of properties. We compare the proposed classifiers against SVM [17] baselines[4] (we used the LibSVM tool [7] in order to select appropriate parameters, which are discussed in each experiment), and against the ensemble approach proposed in [23], called ER (standing for External Referee) in which the ensemble is composed of the base classifiers $\mathcal{C}_{m_1}, \ldots, \mathcal{C}_{m_8}$, but the best classifier for each test instance is selected using a decision tree referee. For associative classifiers, continuous attributes in the training data were discretized using the entropy-minimization method [11], and the attribute-values in the test set were simply mapped to the corresponding intervals (in this way, the discretization process did not use class information in the test set). Experiments that compare classification performance report results for the standard 10-fold cross-validation procedure. In all experiments, parameter k for Algorithm 2 was set to 2 (i.e., each training data, $\mathcal{D}$, was split in two disjoint partitions, d_1 and d_2, in order to obtain the enhanced training data, $\mathcal{D}_e$). In order to ensure that the domain of competence of each metric is independent of a particular data configuration, C^5 and IC^4 were executed a certain number of times employing different data configurations (i.e., training data is randomly partitioned), and the final result is averaged by the number of executions (which was set to 10). To perform a fair comparison, the same procedure was adopted to evaluate the other classifiers, that is, SDC, ER and SVM were also executed multiple times with different data configurations. Best results, including statistical ties, are emphasized. A bold face indicates that the corresponding result was found statistically significant at the 95% confidence level when tested with the two-tailed paired t-test. Experiments were run on 1.8 MHz Intel processors 1GB RAM under Linux.

5.1 Digital Library The first dataset was extracted from the first level of the ACM Computing Classification System (`http://portal.acm.org/dl.cfm/`). The dataset contains 6,682 documents labeled using the 8 first level categories of ACM, namely Hardware (C1), Computer Systems Organization (C2), Software (C3), Computing Methodologies (C4), Mathematics of Computing (C5), Information Systems (C6), Theory of Computation (C7), Computing Milieux (C8). Citations and words in title/abstract compose the set of features. The dataset has a vocabulary of 9,840 unique words, and a total of 51,897 citations. Please, refer to [34] for a detailed description of this dataset.

Using the rules extracted from this dataset, we can analyze the relationship between the widely used confidence metric (m_1) with other metrics, as shown in Figure 1 (to ease the observation of this relationship, we also include, in each graph, a thicker line which indicates the corresponding confidence value). Each point in the graphs corresponds to a rule, for which it is shown the values of some metrics (i.e., confidence in the x-axis and another metric in the y-axis). Clearly, each metric has its particular behavior with varying values of confidence. For lower values of confidence, Added Value (m_2) has a preference for less frequent classes, but, after a certain confidence value, the preference is for more frequent classes. Certainty (m_3) always prefer less frequent classes, but linearly approaches confidence as its value increases. Yules'Q (m_4) and Yules'Y (m_5) have a similar behavior, showing preference for less frequent classes and hardly penalizing associations with low confidence values. Strength Score (m_6) and Weighted Relative Confidence (m_3) both prefer less frequent classes, but Strength Score shows a non-proportional preference for associations with higher values of confidence. The relationship between confidence and support (m_7) is omitted, but, by definition, support shows a preference for more frequent classes. We will use these relationships to explain some of the results reported in the following.

Table 3 shows the classification performance obtained by different classifiers using the ACM dataset (for this application, performance is computed through the traditional accuracy). We will first analyze the performance associated with each category, and then the final classification performance, which is shown in the last line of the table. Classifiers produced by confidence ($\mathcal{C}_{m_1}$) and support ($\mathcal{C}_{m_7}$) performed very well in the most frequent categories (Software, Inf. Systems and Theory of CS). On the other hand, instances belonging to less frequent categories (Comp. Methodologies, Mathematics of CS, and CS Organization) were better classified using Yules'Q ($\mathcal{C}_{m_4}$) and Yules'Y ($\mathcal{C}_{m_5}$). This is expected, and is in agreement with the behaviors depicted in Figure 1 (Yules'Y and Yules'Q show a preference for less frequent categories). The best metric is the one that better balances its performance over all categories. Although the classifier produced by Yules'Y was not the best one for any specific category of ACM, it was the best overall classifier

[4] In most of the experiments, SVM baselines take as input the original training data (and not the enhanced one). However, in some of our experiments, we also employed SVMs as a meta-classifier, as will be detailed.

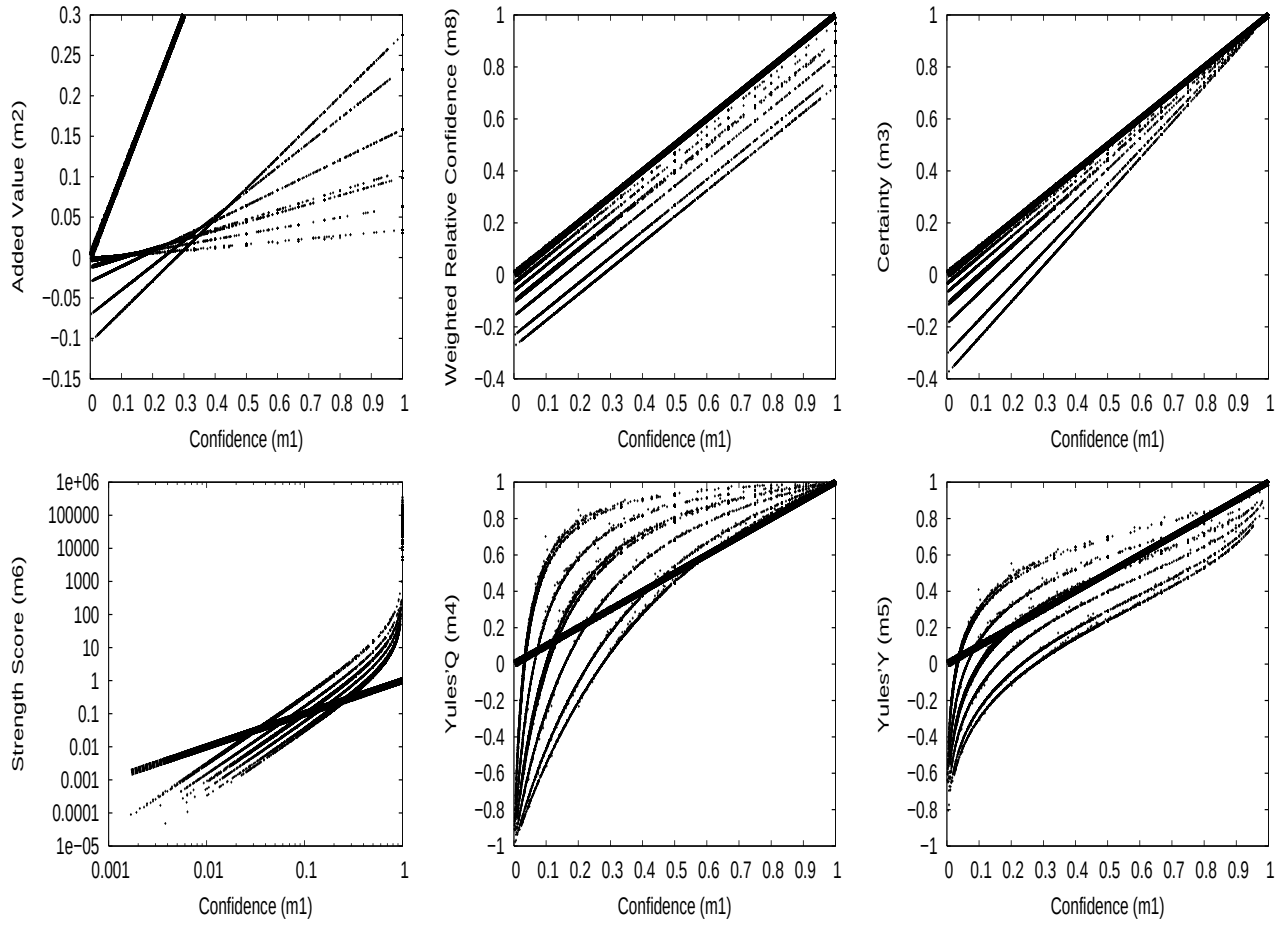

Figure 1: Relationship between confidence and other metrics using the ACM dataset.

	$\mathcal{C}_{m_1}$	$\mathcal{C}_{m_2}$	$\mathcal{C}_{m_3}$	$\mathcal{C}_{m_4}$	$\mathcal{C}_{m_5}$	$\mathcal{C}_{m_6}$	$\mathcal{C}_{m_7}$	$\mathcal{C}_{m_8}$	Div.	Lower Bound	SDC	C^5	IC^4	Upper Bound	ER	SVM
C1	0.808	**0.847**	0.827	0.835	0.833	**0.846**	0.184	0.629	0.235	0.714	0.816	0.809	0.818	0.889	0.799	0.726
C2	0.718	0.790	0.760	0.776	0.798	0.753	0.316	0.789	0.114	0.727	0.727	0.737	0.768	0.878	0.719	**0.877**
C3	0.915	0.852	0.887	0.866	0.862	0.752	**0.960**	0.876	0.124	0.872	0.880	0.889	0.921	0.981	0.875	0.665
C4	0.568	**0.690**	0.626	0.655	0.664	0.678	0.095	0.551	0.228	0.561	0.581	0.618	0.621	0.800	0.605	0.516
C5	0.550	0.625	0.588	0.680	0.678	0.668	0.007	0.325	0.269	0.567	0.570	0.625	0.653	0.750	0.612	**0.909**
C6	**0.952**	0.930	**0.946**	**0.942**	0.926	0.896	0.690	0.766	0.081	0.881	0.919	0.913	0.928	0.968	0.899	0.868
C7	**0.924**	0.894	0.892	0.891	0.891	0.887	0.505	0.682	0.133	0.833	0.902	0.897	0.899	0.924	0.879	0.668
C8	0.643	0.716	0.687	0.722	0.729	**0.756**	0.069	0.478	0.282	0.587	0.658	0.693	0.697	0.824	0.674	**0.770**
Total	0.841	0.847	0.852	0.854	0.854	0.812	0.566	0.733	0.143	0.797	0.849	0.858	**0.882**	0.923	0.812	0.825
Wins	1	1	0	0	0	0	1	0	-	-	0	0	0	-	0	2
Ties	1	1	1	1	0	2	0	0	-	-	0	0	0	-	0	1

Table 3: Classification performance associated with each category of the ACM dataset.

(amongst classifiers produced by other metrics in isolation).

SDC shows a performance that is similar to the performance obtained by most of the base classifiers (the improvement, when it exists, is only marginal). Competence-conscious classifiers C^5 and IC^4 showed the best performances, although they were not the best performer for any individual category. Overall, IC^4 outperformed all other classifiers, providing gains of more than 7%, when compared against SVM (using polynomial kernels of degree 6), and gains of more than 8.5% when compared against ER. IC^4 is always far superior than the corresponding lower bound, but it is also relatively far from the corresponding upper bound.

It seems that the improvements provided by C^5 and IC^4 are somehow related to how accurate and diverse are the base classifiers (we assume that the lower bound is a suitable approximation of the accuracy of the base classifiers, since it tends to increase with the accuracy of the base classifiers). Thus, we relate the harmonic mean between the lower bound and the diversity (i.e., $2 \times \frac{\text{Diversity} \times \text{Lower Bound}}{\text{Diversity+Lower Bound}}$), with the gains relative to the lower bound. Figure 2 shows this relationship. Each point is associated with a category of the ACM dataset. Clearly, for the ACM dataset, the improvements provided by C^5 and IC^4 increase with the diversity and the accuracy of the base classifiers.

We also performed an analysis on how the different metrics were used by C^5 and IC^4, as can be seen in Figure 3 (Left). C^5 utilized only few metrics, specially m_2, m_3 and m_7. Metric m_4 was used to produce classifiers to only one category, and metrics m_5 and m_8 were not used (this is because these two metrics were not the most competent in any category of ACM, and therefore are not considered by C^5). IC^4, on the other hand, utilized all metrics, specially m_1, m_2 and m_3. Both C^5 and IC^4 make large utilization of metrics m_2 and m_3. For C^5, some areas of expertise can be easily detected. Metric m_2 is considered competent for categories Hardware and CS Organization, while metric m_3 is considered competent for category Information Systems. For IC^4, areas of expertise are finer grained, but with manual inspection we detected that m_1 is considered competent for category CS Organization, and m_3 is considered competent for category Milieux.

Next we analyze one of the reasons of the good performance showed by IC^4. Figure 3 (Right) shows the accuracy associated with scenarios for which a different number of metrics are competent. The frequency of occurrence of each scenario is also shown (note that both accuracy and frequency values are shown in the y-axis). As it can be seen, for more than 7% of the instances no metric is competent, and, obviously, these instances were misclassified (this means that the inclusion of other metrics may improve classification performance in this dataset). As expected, accuracy increases with the number of competent metrics. For almost half of the instances all 8 metrics are competent. In these scenarios, there is no risk of misclassification, since a classifier produced by any metric will perform a correct prediction. The accuracy associated with scenarios where only 7 and only 6 of the metrics are competent, is also extremely high (respectively, 99% and 96%). These three scenarios (i.e., 8, only 7, and only 6 metrics are simultaneously competent) correspond to 86% of the instances, and the average accuracy associated with these three scenarios is almost 98% for IC^4. Further, IC^4 shows to be more robust that C^5, providing superior accuracy (relative to the accuracy of C^5) in scenarios where there are only few competent metrics.

We also employed SVMs as meta-classifiers, that is, we used SVMs to select competent metrics for each test instance, and compare the classification performance with the performance obtained by IC^4 and C^5. Our results indicate that the SVM-based meta-learning strategies (i.e., instance-centric and class-centric) lead to classification performances that are (statistically) similar to the classification performances obtained by IC^4 and C^5 counterparts.

5.2 Web Directory Next we investigate the classification performance of competence-conscious associative classifiers using another dataset, which is composed by 2,911 randomly selected articles published in the Slashdot online forum (`http://slashdot.org`). Each article is a document with an author, a title and the story content. Each document is labeled under one of the 16 categories of Slashdot, namely Apple (A1), Ask Slashdot (A2), Backslash (A3), Books (A4), Developers (A5), Entertainment (A6), Games (A7), Hardware (A8), Interviews (A9), Information Technology (A10), Linux (A11), News (A12), Politics (A13), Science (A14), Technology (A15), Your Rights Online (A16). More than a forum for publishing stories, Slashdot constitutes a large social network, where users may interact with each other. Consequently, a particular author may acquire friends and enemies throughout her/his enrollment as a participant of Slashdot. This information about the social relationship of the author (i.e., her/his set of friends and enemies) is explicitly informed in Slashdot and may represent relevant information for the sake of classification.

Classification performance associated with each category is shown in Table 4. Classifiers produced by Strength Score, Support and Relative Confidence achieved the worst performance in most of the categories. SDC did not achieve good performance numbers. IC^4, on the other hand, was the best performer in most of the categories, showing the advantages of competence-conscious classifiers. C^5 achieved a classification performance which was usually close to the performence achieved by ER. In most of the categories, SVM was very competitive with IC^4, but IC^4 was the best overall performer.

Figure 4 shows the relationship between the accuracy and the diversity of the base classifiers, with the improvements provided by C^5 and IC^4. Each point is associated with a category of the Slashdot dataset. As can be seem, the improvements over the lower bound slightly increase with the harmonic mean between the accuracy and the diversity of the base classifiers.

We also investigated the improvements relative to the base classifiers provided by C^5 and IC^4 using simple linear models (a similar approach was used in [25]). Specifically, we are interested in modeling the accuracy of competence-conscious associative classifiers using the best results obtained by the base classifiers. Thus, we assumed a linear relationship between the accuracy obtained by the best base

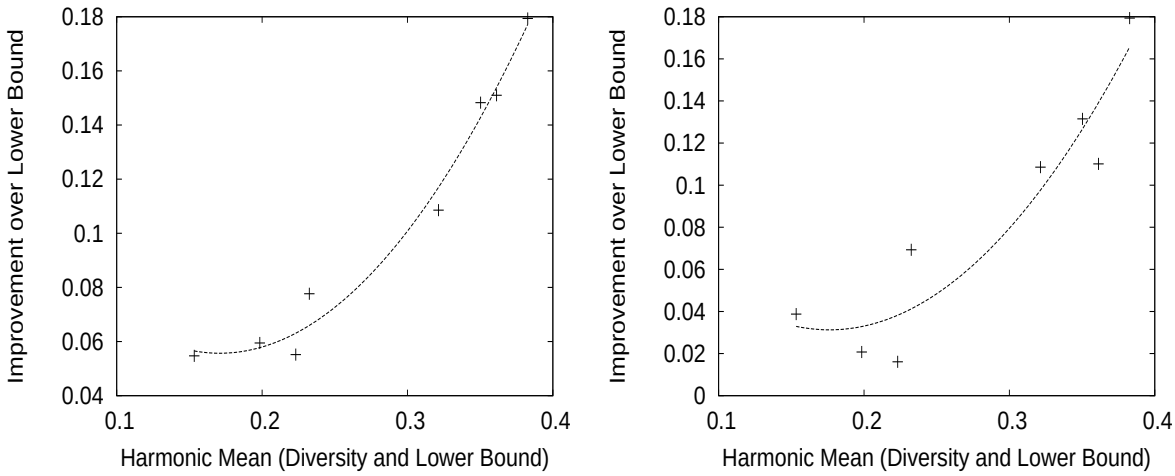

Figure 2: Left − Relationship between diversity and accuracy of the base classifiers, and the improvements provided by IC^4 using the ACM dataset. Right − Relationship between diversity and accuracy of the base classifiers, and the improvements provided by C^5 using the ACM dataset.

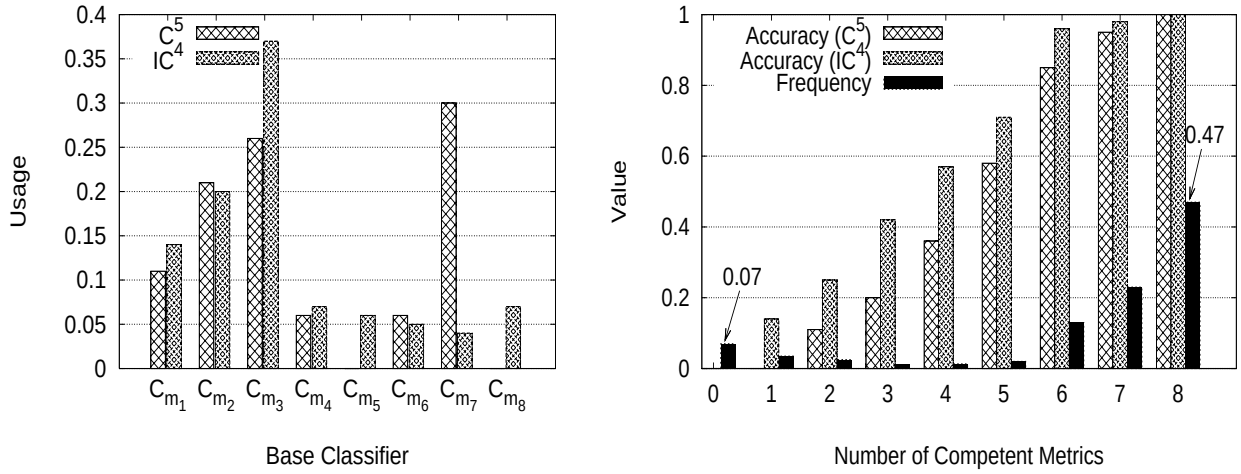

Figure 3: Left − Metric utilization in the ACM dataset. Right − Distribution of competent metrics in the ACM dataset.

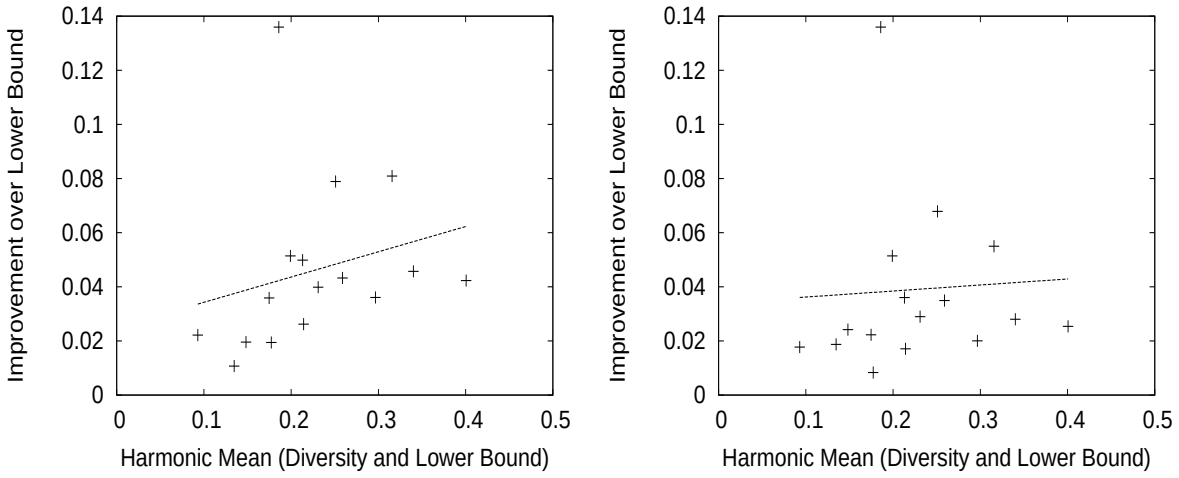

Figure 4: Left – Relationship between diversity and accuracy of the base classifiers, and the improvements provided by IC^4 using the Slashdot dataset. Right – Relationship between diversity and accuracy of the base classifiers, and the improvements provided by C^5 using the Slashdot dataset.

	$\mathcal{C}_{m_1}$	$\mathcal{C}_{m_2}$	$\mathcal{C}_{m_3}$	$\mathcal{C}_{m_4}$	$\mathcal{C}_{m_5}$	$\mathcal{C}_{m_6}$	$\mathcal{C}_{m_7}$	$\mathcal{C}_{m_8}$	Div.	Lower Bound	SDC	C^5	IC^4	Upper Bound	ER	SVM
A1	0.581	0.573	0.596	0.623	0.614	0.600	0.535	0.577	0.168	0.600	0.621	0.622	**0.624**	0.664	0.618	**0.631**
A2	0.886	0.891	0.891	0.893	0.888	0.862	0.849	0.870	0.121	0.882	0.889	0.892	**0.903**	0.925	0.896	**0.899**
A3	0.723	0.704	0.723	0.703	0.700	0.689	0.659	0.678	0.097	0.723	0.721	0.732	**0.738**	0.801	**0.734**	**0.736**
A4	0.481	0.489	0.500	0.503	0.508	0.472	0.434	0.474	0.214	0.498	0.501	**0.510**	**0.515**	0.679	**0.513**	**0.513**
A5	0.591	0.590	0.591	0.610	0.598	0.547	0.538	0.558	0.306	0.595	0.604	0.606	**0.612**	0.683	0.601	**0.611**
A6	0.550	0.547	0.575	0.560	0.568	0.542	0.522	0.551	0.143	0.556	0.558	**0.570**	**0.577**	0.684	**0.572**	**0.567**
A7	**0.929**	0.901	0.906	0.915	0.920	0.873	0.845	0.877	0.103	0.805	0.912	0.921	0.916	0.975	0.917	0.900
A8	0.843	0.856	0.870	0.871	0.870	0.837	0.815	0.836	0.080	0.864	0.889	**0.889**	0.885	0.941	0.885	**0.894**
A9	0.801	0.806	0.805	0.803	0.809	0.783	0.781	0.778	0.099	0.806	0.811	0.823	**0.836**	0.896	0.828	0.825
A10	0.525	0.521	0.550	0.568	0.565	0.508	0.500	0.506	0.167	0.547	0.558	**0.582**	**0.590**	0.683	**0.582**	**0.580**
A11	0.901	0.909	0.909	0.908	0.905	0.886	0.877	0.886	0.047	0.906	**0.916**	**0.918**	**0.918**	0.983	**0.913**	0.902
A12	0.630	0.633	0.646	0.639	0.645	0.623	0.619	0.629	0.212	0.622	0.649	0.651	**0.669**	0.707	0.653	**0.671**
A13	0.680	0.685	0.689	0.688	0.691	0.658	0.684	0.665	0.230	0.680	0.692	0.695	**0.708**	0.791	0.701	0.696
A14	0.865	0.854	0.867	0.863	0.860	0.856	0.846	0.849	0.114	0.837	0.865	**0.875**	**0.877**	0.934	**0.872**	**0.876**
A15	0.748	0.757	0.761	0.751	0.753	0.742	0.744	0.746	0.077	0.750	**0.763**	**0.761**	**0.757**	0.802	**0.765**	**0.757**
A16	0.733	0.735	0.735	0.731	0.725	0.730	0.701	0.726	0.124	0.722	0.737	0.747	**0.762**	0.829	**0.760**	0.748
Total	0.729	0.721	0.756	0.765	0.762	0.700	0.684	0.704	0.220	0.735	0.765	0.785	**0.827**	0.964	0.791	0.815
Wins	1	0	0	0	0	0	0	0	-	-	0	0	2	-	0	0
Ties	0	0	0	0	0	0	0	0	-	-	2	7	12	-	8	11

Table 4: Classification performance associated with each category of the Slashdot dataset.

classifier in each category, and the corresponding accuracy obtained by either C^5 or IC^4. We characterized this relation using statistical correlation coefficients (CC).

The associated regression lines were built using the 16 categories of Slashdot (i.e., each point corresponds to one of the 16 categories). Regression lines for C^5 and IC^4 are shown in Figure 5, and both have very high correlation coefficients (which are shown between parenthesis). Further, their regression gradients are higher than one, possibly indicating that, in the limit, competence-conscious associative classifiers are indeed more accurate than the best base associative classifier.

5.3 Web Spam Detection In this application the objective is to detect malicious actions aimed at the ranking functions used by search engines. We used a dataset obtained from the Web Spam Challenge (`http://webspam.lip6.fr/wiki/pmwiki.php`). The dataset is very skewed (only 6% of the examples are spam pages). Each example is composed of direct features (i.e., number of pages in the host, number of characters in the host name etc.) link-based features (i.e., in-degree, out-degree, PageRank etc.) and content-based features (i.e., number of words in the page, average word length etc.).

Table 5 shows the classification performance obtained

	$\mathcal{C}_{m_1}$	$\mathcal{C}_{m_2}$	$\mathcal{C}_{m_3}$	$\mathcal{C}_{m_4}$	$\mathcal{C}_{m_5}$	$\mathcal{C}_{m_6}$	$\mathcal{C}_{m_7}$	$\mathcal{C}_{m_8}$	Lower Bound	SDC	C^5	IC^4	Upper Bound	ER	SVM
Acc.	**0.945**	0.705	0.703	0.895	0.900	**0.946**	**0.945**	0.881	0.854	0.861	0.872	0.897	0.988	0.864	**0.954**
F1	0.484	0.525	0.525	0.585	0.592	0.591	0.487	0.589	0.590	0.593	0.611	**0.622**	0.946	0.587	0.502
AUC	0.500	0.755	0.758	0.608	0.605	0.559	0.498	0.632	0.662	0.727	0.715	**0.790**	0.911	0.728	0.514

Table 5: Classification performance for Web spam detection.

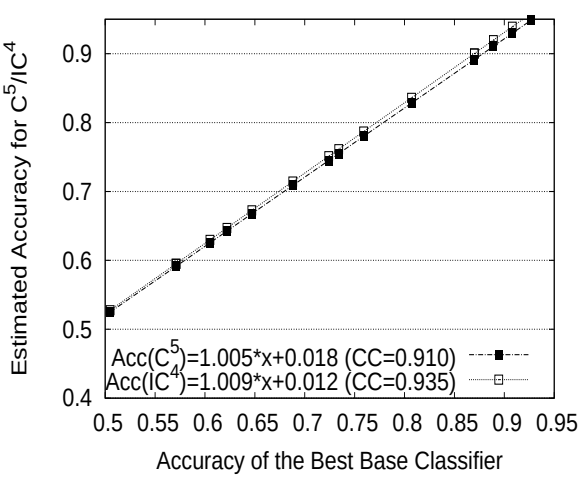

Figure 5: Accuracy model for competence-conscious associative classifiers using the Slashdot dataset.

by different classifiers (for this application, performance is computed through accuracy, F_1 measure[5], and the area under the curve). $\mathcal{C}_{m_1}$ and $\mathcal{C}_{m_7}$ showed impressive performance in terms of accuracy. This is expected, because the vast majority of examples are legitimate pages, and confidence (m_1) and support (m_7) have preference for more frequent classes. On the other hand, $\mathcal{C}_{m_1}$ and $\mathcal{C}_{m_7}$ showed poor performance in terms of F_1 and AUC (i.e., no spam pages were actually detected). The remaining base associative classifiers were able to detect some spam pages, specially $\mathcal{C}_{m_6}$, which also shows impressive performance in terms of accuracy. In terms of AUC, $\mathcal{C}_{m_2}$ and $\mathcal{C}_{m_3}$ showed the best performance, amongst base classifiers. Thus, different metrics show distinct performance depending to the evaluation target (i.e., accuracy, F_1 or AUC).

Now we evaluate C^5 and IC^4, which are the best performers in terms of F_1. Although IC^4 showed to be far from the optimal performance, it showed impressive gains when compared against SVM (using linear kernels with parameter C set to 5.00), and ER, in terms of F_1 and AUC.

Results obtained by detection methods that were specifically designed for Web spam detection are shown in `http://webspam.lip6.fr/wiki/pmwiki.php?n=Main.PhaseIIIResults`. These methods use specific heuristics that are particularly suitable for Web spam detection. The first six competitors achieved detection performances that range from 0.731 to 0.848 in terms of AUC. IC^4 shows to be very competitive, achieving a detection performance which is (statistically) similar to the performance obtained by the 4th place.

5.4 UCI Datasets In the last set of experiments, we used 26 datasets obtained from the UCI Machine Learning Repository [5]. Table 7 shows the performance obtained by each classifier using these datasets (for this application, performance is computed through the traditional accuracy). $\mathcal{C}_{m_4}$ and $\mathcal{C}_{m_5}$ showed poor performance in skewed datasets where few classes are much more frequent than the others (i.e., anneal, lymph, auto, hypo). This is because Yules'Y and Yules'Q have preference for less frequent classes (as shown in Figure 1). For these skewed datasets, $\mathcal{C}_{m_7}$ (support) showed its best performance, since the likelihood of predicting most frequent classes is higher in such datasets (this is expected, due to the definition of support). For most of the datasets, $\mathcal{C}_{m_1}$, $\mathcal{C}_{m_2}$ and $\mathcal{C}_{m_3}$ are in close rivalry ($\mathcal{C}_{m_1}$ shows a slightly better average performance, but $\mathcal{C}_{m_3}$ shows better performance more often). $\mathcal{C}_{m_6}$ (strength score) shows the best average performance, amongst all base classifiers.

On average, C^5 shows superior classification performance than SDC. Also, the performance of C^5 is, on average, slightly superior than the performance obtained by SVM (parameters for each dataset are shown in Table 6) and ER baselines. Again, IC^4 is the best performer, and for some datasets it reaches a performance that is close to optimal (i.e., anneal, breast, hypo, iris, labor, sick, wave and wine), suggesting that the more fine-grained the analysis of competence, the more effectively the metrics are combined. Interestingly, the performance of C^5 approaches the performance of IC^4 for datasets containing more classes (i.e., glass, led7, lymph, vehicle, and zoo), since in this case the competence analysis performed by C^5 becomes finer grained.

Some datasets deserve special attention. IC^4 showed very good performance in the anneal dataset. Figure 6(left) shows the frequency distribution of competent metrics for this dataset. Almost 70% of the instances have more than five competent metrics, and in such scenarios accuracy reaches 100%. The accuracy obtained in such scenarios guarantees a final classification performance that is already superior

[5]A combination of precision (p) and recall (r) defined as their harmonic mean $\frac{2pr}{p+r}$.

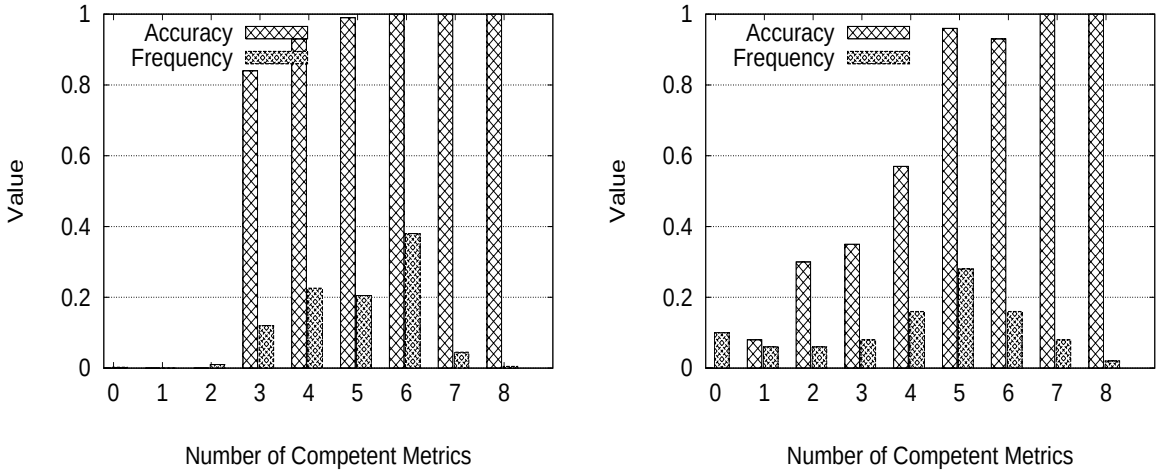

Figure 6: Distribution of Competent Metrics for IC^4 in the anneal (left) , and auto (right) Datasets.

	Kernel	C	degree	γ
anneal	polynomial	-	6	-
austra	linear	1.50	-	-
auto	RBF	-	-	0.0003
breast	linear	3.00	-	-
cleve	linear	3.00	-	-
crx	linear	0.10	-	-
diabet	linear	1.00	-	-
germa	linear	1.00	-	-
glass	RBF	-	-	0.0012
heart	linear	5.00	-	-
hepati	linear	0.10	-	-
horse	linear	1.00	-	-
hypo	linear	3.00	-	-
iono	linear	0.50	-	-
iris	polynomial	-	4	-
labor	linear	0.50	-	-
led7	RBF	-	-	0.0012
lymph	polynomial	-	5	-
pima	linear	0.10	-	-
sick	linear	5.00	-	-
sonar	linear	5.00	-	-
tic-tac	linear	0.50	-	-
vehicle	polynomial	-	6	-
wave	polynomial	-	5	-
wine	polynomial	-	5	-
zoo	RBF	-	-	0.0012

Table 6: SVM parameters for each UCI dataset.

than the performance of $\mathcal{C}_{m_2}$, $\mathcal{C}_{m_4}$, $\mathcal{C}_{m_5}$ and $\mathcal{C}_{m_8}$. Similar trends also happens in datasets austra, breast, cleve, german, heart, hypo, iono, iris, sick and wine. In the auto dataset IC^4 showed poor performance, being worse than base classifiers $\mathcal{C}_{m_1}$, $\mathcal{C}_{m_2}$, $\mathcal{C}_{m_3}$ and $\mathcal{C}_{m_6}$. Figure 6(right) shows the frequency/accuracy distribution of competent metrics for this dataset. As can be seem, the accuracy associated with almost 40% of the instances falls below 58%, which are the scenarios with less than 5 competent metrics. Also, we believe that, for such datasets, the meta-classifier was not able to correctly distinguish the domains of competence. Similar trend also happens for datasets hepati, tic-tac, and wave.

Figure 7 shows the relationship between the diversity and the accuracy (i.e., the lower bound) of the base classifiers (i.e., $2 \times \frac{\text{Diversity} \times \text{Lower Bound}}{\text{Diversity+Lower Bound}}$) and the improvements provided by C^5 and IC^4 over the lower bound. Each point is associated with an UCI dataset. In general, the improvements increase with the harmonic mean between the diversity and the accuracy of the base classifiers. This suggests the importance of the diversity and the accuracy of the base classifiers, in order to produce effective ensembles.

We finish our evaluation with simple linear models that are used to assess the improvements provided by C^5 and IC^4, relative to the base classifiers. Again, we are interested in modeling the accuracy of competence-conscious associative classifiers using the best results obtained by the base classifiers.

The associated regression lines were built using the 26 UCI datasets (i.e., each point corresponds to one of the 26 datasets). Regression lines for C^5 and IC^4 are shown in Figure 8, and both have very high correlation coefficients. Further, their regression gradients are higher than one (note that this makes the estimated accuracy higher than one when x=1.00), possibly indicating that, in the limit, competence-conscious associative classifiers are indeed more accurate than the best base associative classifier.

We also employed SVMs as meta-classifiers, that is, we used SVMs to select competent metrics for each test instance, and compare the classification performance with the performance obtained by IC^4 and C^5. The results we

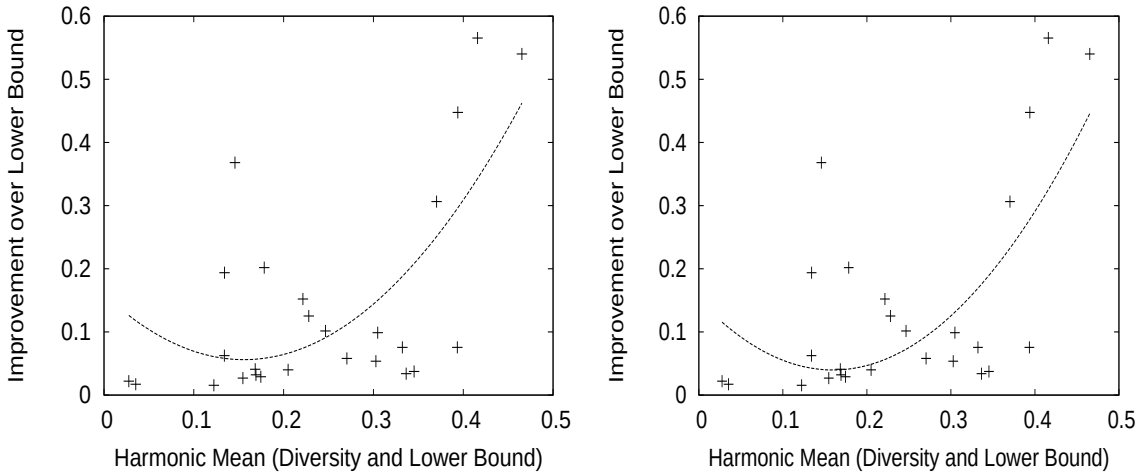

Figure 7: Left − Relationship between diversity and accuracy of the base classifiers, and the improvements provided by IC^4 using the UCI datasets. Right − Relationship between diversity and accuracy of the base classifiers, and the improvements provided by C^5 using the UCI datasets.

	$\mathcal{C}_{m_1}$	$\mathcal{C}_{m_2}$	$\mathcal{C}_{m_3}$	$\mathcal{C}_{m_4}$	$\mathcal{C}_{m_5}$	$\mathcal{C}_{m_6}$	$\mathcal{C}_{m_7}$	$\mathcal{C}_{m_8}$	Div.	LB	SDC	C^5	IC^4	UB	ER	SVM
anneal	0.761	0.694	0.864	0.115	0.229	0.926	0.763	0.616	0.374	0.623	0.926	0.935	**0.958**	0.993	0.921	0.946
austra	0.861	0.855	0.852	0.861	0.849	0.856	0.833	0.864	0.084	0.859	**0.869**	**0.868**	**0.880**	0.921	0.848	0.855
auto	0.715	0.752	0.759	0.043	0.105	**0.782**	0.404	0.513	0.280	0.534	0.684	0.705	0.698	0.895	0.674	0.725
breast	0.940	**0.971**	**0.969**	**0.967**	**0.972**	**0.971**	0.934	**0.972**	0.017	0.950	0.955	**0.961**	**0.970**	0.989	**0.973**	**0.973**
cleve	**0.843**	0.832	0.826	0.839	0.832	0.816	0.835	0.824	0.069	0.839	**0.842**	0.836	**0.853**	0.900	**0.839**	0.836
crx	0.835	0.850	0.853	0.849	0.841	**0.860**	0.843	**0.864**	0.091	0.847	0.855	**0.865**	**0.873**	0.926	**0.862**	**0.855**
diabet	**0.783**	0.745	0.739	0.749	0.752	**0.779**	0.698	0.738	0.188	0.741	0.750	**0.783**	**0.782**	0.937	0.776	0.766
germa	0.702	0.694	0.693	0.692	0.693	**0.748**	0.698	0.723	0.226	0.717	0.727	**0.734**	**0.748**	0.953	**0.738**	0.712
glass	**0.713**	0.657	0.672	0.640	0.648	**0.709**	0.565	0.646	0.286	0.633	0.657	**0.701**	0.682	0.865	0.672	**0.705**
heart	0.814	0.840	0.835	0.822	0.830	0.829	0.827	0.830	0.121	0.828	0.832	0.840	**0.861**	0.901	**0.848**	0.838
hepati	0.798	0.778	0.781	0.841	0.829	0.848	0.797	**0.850**	0.161	0.797	0.805	0.837	0.835	0.989	0.818	0.810
horse	0.710	0.731	0.732	0.702	0.685	0.747	0.773	0.720	0.192	0.743	0.756	0.771	**0.813**	0.890	0.777	**0.822**
hypo	0.951	0.879	0.881	0.126	0.129	0.971	0.957	0.932	0.084	0.730	0.883	0.937	**0.992**	1.000	0.953	0.988
iono	0.901	0.896	0.888	0.875	0.870	**0.929**	0.688	0.840	0.145	0.858	0.898	0.910	**0.942**	0.980	0.918	0.917
iris	**0.940**	**0.950**	**0.948**	**0.946**	**0.951**	0.936	**0.944**	**0.942**	0.021	0.937	**0.948**	**0.947**	**0.949**	0.956	**0.947**	**0.959**
labor	**1.000**	0.948	0.930	0.758	0.890	0.951	0.627	0.924	0.095	0.965	0.969	**1.000**	**0.997**	1.000	0.973	0.782
led7	0.740	0.741	0.739	0.737	0.743	0.711	0.743	0.741	0.222	0.733	0.746	**0.774**	**0.758**	0.806	0.745	0.748
lymph	**0.862**	0.758	0.812	0.064	0.143	0.780	0.747	0.781	0.299	0.579	0.798	**0.844**	**0.846**	0.948	**0.849**	0.802
pima	0.733	0.746	0.748	0.749	0.749	0.777	0.692	0.745	0.212	0.740	0.750	0.770	**0.798**	0.936	0.771	0.768
sick	0.936	0.640	0.648	0.128	0.139	0.967	0.934	0.677	0.315	0.623	0.928	0.946	**0.983**	0.997	0.969	0.965
sonar	0.813	**0.866**	**0.862**	0.854	**0.865**	0.834	0.770	**0.864**	0.072	0.821	0.855	**0.867**	**0.868**	0.952	**0.865**	0.841
tic-tac	0.649	**0.918**	0.814	**0.928**	**0.927**	0.816	0.410	0.530	0.105	0.766	0.836	0.880	**0.926**	1.000	0.876	0.830
vehicle	0.658	0.670	0.664	0.668	0.648	0.701	0.534	0.623	0.134	0.655	0.665	**0.728**	**0.734**	0.765	0.700	**0.722**
wave	0.809	0.808	0.807	0.814	0.813	0.813	0.787	0.788	0.097	0.802	0.814	0.814	0.834	0.836	0.821	**0.871**
wine	0.913	0.934	0.932	0.818	0.826	**1.000**	0.687	0.745	0.068	0.831	0.885	**1.000**	**0.990**	1.000	0.941	0.975
zoo	0.845	0.849	0.876	0.856	0.781	0.912	0.708	0.698	0.133	0.826	0.894	**0.942**	**0.952**	0.973	0.956	0.928
Avg.	0.816	0.808	0.812	0.671	0.682	0.845	0.738	0.769	0.157	0.768	0.828	**0.854**	**0.866**	0.935	0.847	0.844
Wins	0	0	0	0	0	1	0	1	-	-	0	0	4	-	0	1
Ties	6	4	3	3	4	7	1	4	-	-	3	14	19	-	8	6

Table 7: Classification performance of classifiers in the UCI datasets.

obtained indicate that, again, the SVM-based meta-learning strategies (i.e., instance-centric and class-centric) lead to classification performances that are (statistically) similar to the classification performances obtained by IC^4 and C^5 counterparts.

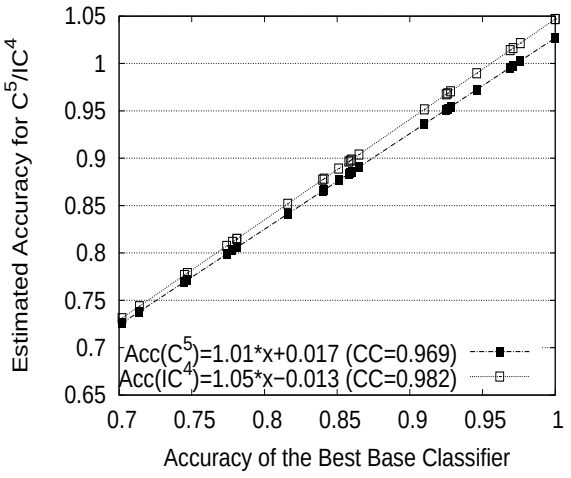

Figure 8: Accuracy model for competence-conscious associative classifiers using the UCI datasets.

6 Conclusions

This paper focused on an important problem in associative classification, the *metric dilemma*. We have shown that the performance of associative classifiers are strongly dependent on the metric that is used to quantify the strenght of the association between features and classes. There is no perfect metric, and no metric is consistently superior than all others, in the sense that it can be safely used in isolation. In fact, each metric has a particular domain of competence, or area of expertise, for which it is able to produce the most accurate classifier. We investigate meta-learning methods, which use the training data to learn the domain of competence of each metric. Finally, the competence of metrics are exploited to decide which is the best metric to be applied in each scenario, resulting in a combination or ensemble of classifiers produced by different metrics, which maximizes the performance of the final classifier, that we denoted as *competence-conscious* associative classifiers.

For effective metric combination, the corresponding classifiers must cover different portions of the training data (i.e., the metrics must show some diversity), and the training data must have features that are able to distinguish those portions of the training data. If these favorable conditions are met, our method reaches full potential of the base classifiers (i.e., the performance is close to the upper bound). On the other hand, a performance penalty may result.

We proposed competence-conscious classifiers, where the difference between them resides in how they perform the analysis of the domains of competence. The coarse-grained analysis, performed by the class-centric approach (C^5) provides lower gains when compared to the fine-grained analysis, performed by the instance-centric approach (IC^4), which outperforms all the evaluated classifiers, including a simple delegation approach (SDC), an existing ensemble method (ER), and SVMs. As future work, we intend to move forward by investigating other application scenarios and evaluating other association metrics. We are currently studying the correlation between metrics and the effectiveness of the corresponding associative classifiers.

Acknowledgements

This research is based upon work supported by UOL (www.uol.com.br) through its UOL Bolsa Pesquisa program (grant number 20080131200100). This work is also supported in part by CNPq, Capes, Finep, Fapemig (project number 14281), and by the projects 5S-VQ (CNPq grant number 551013/2005-2), INCTWeb (CNPq grant number 573871/2008-6), and InfoWeb (CNPq grant number 550874/2007-0).

References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *SIGMOD*, pages 207–216. ACM, 1993.

[2] M. Antonie, O. Zaïane, and R. Holte. Learning to use a learned model: A two-stage approach to classification. In *ICDM*, pages 33–42, 2006.

[3] B. Arunasalam and S. Chawla. CCCS: a top-down associative classifier for imbalanced class distribution. In *KDD*, pages 517–522. ACM, 2006.

[4] E. Baralis and P. Garza. Majority classification by means of association rules. In *PKDD*, pages 35–46, 2003.

[5] C. Blake and C. Merz. UCI repository of machine learning datasets. 1998.

[6] L. Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, 1996.

[7] C.-C. Chang and C.-J. Lin. *LibSVM: A Library for Support Vector Machines*, 2001. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[8] H. Cheng, X. Yan, J. Han, and C. Hsu. Discriminative frequent pattern analysis for effective classification. In *ICDE*, pages 716–725, 2007.

[9] H. Cheng, X. Yan, J. Han, and P. Yu. Direct discriminative pattern mining for effective classification. In *ICDE*, pages 169–178, 2008.

[10] W. Fan, K. Zhang, H. Cheng, J. Gao, X. Yan, J. Han, P. Yu, and O. Verscheure. Direct mining of discriminative and essential frequent patterns via model-based search tree. In *KDD*, pages 230–238, 2008.

[11] U. Fayyad and K. Irani. Multi interval discretization of continuous-valued attributes for classification learning. In *IJCAI*, pages 1022–1027. M. Kaufmann, 1993.

[12] C. Ferri, P. Flach, and J. Hernández-Orallo. Delegating classifiers. In *ICML*, page 37. ACM, 2004.

[13] Y. Freund. Boosting a weak learning algorithm by majority. *Inf. Comput.*, 121(2):256–285, 1995.

[14] J. Friedman. Comment on "Classifier Technology and the

Illusion of Progress" by D. Hand. *Statistical Science*, 21(1):15–18, 2006.
[15] J. Gama and P. Brazdil. Cascade generalization. *Machine Learning*, 45:315–343, 2000.
[16] R. Hilderman and H. Hamilton. Evaluation of interestingness measures for ranking discovered knowledge. In *PAKDD*, pages 247–259. Springer, 2001.
[17] T. Joachims. Training linear SVMs in linear time. In *KDD*, pages 217–226. ACM, 2006.
[18] A. Juditsky, A. Nazin, A. Tsybakov, and N. Vayatis. Recursive aggregation of estimators by the mirror descent algorithm with averaging. *Probl. Inf. Transm.*, 41(4):368–384, 2005.
[19] S. Kotsiantis and P. Pintelas. Bagged voting ensembles. In *AIMSA*, pages 168–177, 2004.
[20] N. Lavrac, P. Flach, and B. Zupan. Rule evaluation measures: A unifying view. *Inductive Logic Prog.*, 1634:174–185, 1999.
[21] W. Li, J. Han, and J. Pei. CMAR: Accurate and efficient classification based on multiple class-association rules. In *ICDM*, pages 369–376. IEEE, 2001.
[22] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *KDD*, pages 80–86. ACM, 1998.
[23] J. Ortega, M. Koppel, and S. Argamon. Arbitrating among competing classifiers using learned referees. *KAIS*, 3:470–490, 2001.
[24] R. Schapire. A brief introduction to boosting. In *IJCAI*, pages 1401–1406. M. Kaufmann, 1999.
[25] A. Seewald. Exploring the parameter state space of stacking. In *ICDM*, pages 685–688, 2002.
[26] Y. Sun, Y. Wang, and A. Wong. Boosting an associative classifier. *IEEE Trans. Knowl. Data Eng.*, 18(7):988–992, 2006.
[27] Y. Sun, A. Wong, and Y. Wang. An overview of associative classifiers. In *DMIN*, pages 138–143, 2006.
[28] P. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *KDD*, pages 32–41. ACM, 2002.
[29] R. Thonangi and V. Pudi. Acme: An associative classifier based on maximum entropy principle. In *ALT*, pages 122–134, 2005.
[30] K. Ting and I. Witten. Issues in stacked generalization. *Journal of Artificial Intelligence Res.*, (10):271–289, 1999.
[31] A. Tsymbal, M. Pechenizkiy, and P. Cunningham. Diversity in random subspacing ensembles. In *DaWaK*, pages 309–319, 2004.
[32] A. Tsymbal, M. Pechenizkiy, and P. Cunningham. Dynamic integration with random forests. In *ECML*, pages 801–808, 2006.
[33] A. Veloso, W. M. Jr., and M. J. Zaki. Lazy associative classification. In *ICDM*, pages 645–654. IEEE, 2006.
[34] A. Veloso, W. Meira, M. Cristo, M. Gonçalves, and M. Zaki. Multi-evidence, multi-criteria, lazy associative document classification. In *CIKM*, pages 218–227. ACM, 2006.
[35] A. Veloso, W. Meira, M. Gonçalves, and M. Zaki. Multi-label lazy associative classification. In *PKDD*, pages 605–612, 2007.
[36] A. Veloso, H. Mosrri, M. Gonçalves, and W. Meira. Learning to rank at query-time using association rules. In *SIGIR*, pages 267–274. ACM, 2008.
[37] J. Wang and G. Karypis. HARMONY:efficiently mining the best rules for classification. In *SDM*. SIAM, 2005.
[38] D. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.
[39] Y. Yoon and G. Lee. Efficient implementation of associative classifiers for document classification. *Inf. Process. Manage.*, 43(2), 2007.
[40] G. Yule. On the association of attributes in statistics. *Phil. Trans. A*, (194):257–319, 1900.