Pattern Recognition – Applications and Methods

International Conference, ICPRAM 2012 Vilamoura, Algarve, Portugal, February 6-8, 2012
Revised Selected Papers
Pedro Latorre Carmona, J. Salvador Sánchez, and Ana L.N. Fred (Eds.)

Pattern Recognition –
Applications
and Methods

International Conference, ICPRAM 2012
Vilamoura, Algarve, Portugal,
February 6–8, 2012
Revised Selected Papers

Springer
Adaptive Features for Object Classification .......................... 121
Heydar Maboudi Afkham, Stefan Carlsson, Josephine Sullivan

SVM-Based Feature Selection and Classification for Email Filtering .... 135
Sebastián Maldonado, Gaston L’Huillier

Instance Selection Methods and Resampling Techniques for Dissimilarity
Representation with Imbalanced Data Sets ............................. 149
M. Millán-Giraldo, V. García, J.S. Sánchez

Simplex Decompositions Using Singular Values Decomposition ........ 161
Madhusudana Shashanka, Michael Giering

On the Understanding of the Stream Volume Behavior on Twitter .... 171
Juan Zamora, Miguel Guevara, Gabriela Domínguez, Héctor Allende,
Alejandro Veloz, Rodrigo Salas

Training Selection with Label Propagation for Semi-supervised Land
Classification and Segmentation of Satellite Images .................. 181
Olga Rajadell, Pedro García-Sevilla

Hyperspectral Imagery Framework for Unmixing and Dimensionality
Estimation ................................................................. 193
José M.P. Nascimento, José M. Bioucas-Dias

Author Index ............................................................. 205
Abstract. In the dissimilarity representation approach, the dimension reduction of the dissimilarity space is addressed by using instance selection methods. Several studies have shown that these methods work well on small data sets. Also, the uniformity of the instances distribution can be obtained when the classes are evenly spread and balanced. However, many real-world problems are characterized by an imbalanced class distribution. In this paper, we address the problem of instance selection for constructing the dissimilarity space in the imbalanced data context. Class imbalance is handled by resampling the data set, whereas instance selection is applied to find a small representation set. Experimental results demonstrate the significance of the joint use of resampling techniques and instance selection methods to improve the performance of classifiers trained on dissimilarity representation.

Keywords: Instance selection, Dissimilarity representation, Resampling techniques, Imbalanced data sets.

1 Introduction

The statistical pattern recognition approach traditionally represents the objects in vector spaces by a set of measurable features. However, this approach presents some drawbacks: (i) objects of different classes may be represented by the same feature vectors and (ii) the classifiers could be affected by the variation of feature sets [1]. An alternative approach to the feature-based representation that overcomes these problems is the dissimilarity representation paradigm proposed by Pekaslka and Duin [2]. Here, the objects are represented by their dissimilarity or distance values to the other objects in the set.

The construction of a new vector space from a dissimilarity representation is carried out in two ways [3]: (i) Euclidean embedding and (ii) the dissimilarity space. The former case is based on embedding the given non-Euclidean dissimilarity data into a vector space preserving the distances between objects as good as possible in comparison to the original dissimilarities. The second way postulates an Euclidean vector space defined by the dissimilarities vectors. This method considers the dissimilarity matrix...
as a new training data set, where the set of rows (dissimilarities) vectors (one for each object) represents individual training samples and the columns form the dimensions of the so-called dissimilarity space. For its construction, the pairwise dissimilarities are computed between a given object and objects from the representation set $R$. In general, a representation set is a set of chosen prototypes of the training set $T$. Sometimes, $R$ can be chosen as the whole training set.

In the dissimilarity space, the dimensionality is determined by the size of the representation set. When all training objects are used to build the representation set, the dimension of the dissimilarity space is equal to $|T|$, which may impose a computational burden on the classifier. To overcome this problem, numerous works have proposed to use and develop instance selection methods (instance selection methods) for finding a small reduced representation set (from the training data) capable of achieving a good trade-off between classification accuracy and computational efficiency [4,5,6,7,8,9,10]. Results using instance selection methods have shown a good performance for small training sets. Likewise, when the classes are evenly spread and balanced, it is possible to gain a uniform prototypes distribution. However, in many real-world problems, there exists an extremely skewed difference between the class ratios of prior probabilities. This data complexity, known as the class imbalance problem [11], may affect the instance selection process to obtain reduced representation sets that does not reflect the true distribution [9].

The class imbalance problem occurs when one class vastly outnumbers the other class, which is usually the most important one and with the highest misclassification costs. Instances from the minority and majority classes are often referred to as positive and negative, respectively. Several solutions have been proposed to deal with this data complexity. One of the most investigated is resampling, which aims at balancing the original data set, either by over-sampling the minority class [12,13] and/or by under-sampling the majority class [14,15], until the classes are approximately equally represented.

Although class imbalance has been extensively studied for binary classification problems, very few approaches explore the class imbalance problem in the dissimilarity space [16,17,18]. Besides, to the best of our knowledge, no work has been carried out on how to select a small representation set for constructing the dissimilarity space on imbalanced data sets.

This paper investigates some strategies to select a reduced representation set and manage the class imbalance for dissimilarity representation. In order to face such a problem, this work focuses on the joint use of instance selection methods and resampling techniques. To this end, we will carry out experiments over real data sets, employing four renowned instance selection methods and two resampling algorithms. All techniques are evaluated in terms of their geometric mean of accuracies, and then compared for statistical differences using the Friedman’s average rank test and the Nemenyi’s post hoc test.

The rest of the paper is outlined as follows. Section 2 provides a summary of the classification problem in dissimilarity representation. Section 3 presents a brief overview of instance selection methods. An introduction to resampling algorithms is provided
in Section 4. In Section 5, the experimental setup is described. Next, in Section 6, the results are showed and discussed. Finally, Section 7 concludes the present study.

2 Dissimilarity Space

In traditional pattern recognition algorithms, objects are represented by a vector of features, in which the dimensionality of the feature space is given by the number of features employed to describe the objects. On the contrary, in the dissimilarity space, objects are represented by dissimilarity vectors, where each element of a vector relates an object with other objects [2].

Given a training set of \( n \) objects, \( T = \{x_1, \ldots, x_n\} \), a new set of \( r \) representative objects of the problem, called prototypes, is obtained from \( T \). This set of prototypes, which contains information of all classes in \( T \), is known as representation set, \( R = \{p_1, \ldots, p_r\} \). The amount of prototypes (\( r \)) in \( R \) determines the dimension of the dissimilarity space. Several methods have been proposed in the literature to select this set of prototypes; for example, Pekalska et al. studied the random and systematic selection procedures for the normal density-based quadratic classifier [8].

In dissimilarity-based classification, some dissimilarity measure \( d \) has to be employed to compute the proximity between objects. Given the pair of objects \( x = (x_1, x_2, \ldots, x_m) \) and \( y = (y_1, y_2, \ldots, y_m) \), the dissimilarity measure \( d \) must satisfy one or more of the usual conditions for a metric [19]: non-negativity, identity of indiscernibles, symmetry and triangle inequality.

Usually, the dissimilarity measure used to represent objects by proximities corresponds to the Euclidean distance between two objects \( x \) and \( y \), that is,

\[
d(x, y) = \left( \sum_{j=1}^{m} (x_j - y_j)^2 \right)^{1/2}
\]

where \( m \) is the number of features. Then, the proximity between the \( i \)-th object in \( T \), \( x_i \), and all prototypes in \( R \) is

\[
D(x_i, R) = \{d(x_i, p_1), \ldots, d(x_i, p_r)\}
\]

which is a vector with \( r \) distances that associates \( x_i \) with all objects in \( R \). By doing \( D(T, R) \), a \( n \times r \) dissimilarity matrix is obtained, which refers to the distances from the objects in the training set to all objects in the representation set and it will be further used to built the classifier in the dissimilarity space.

In this paper, we will use the Euclidean distance measure. Given a test set \( S \), the proximity between objects in \( S \) and prototypes in \( R \) is also computed, giving a dissimilarity matrix \( D(R, S) \). Thus, the test set \( S \) can be evaluated with the classifier built in the dissimilarity space.

3 Instance Selection Methods

In the framework of the dissimilarity representation, the instance selection methods are used to find a small representation set for reducing the computational effort, while
preserving the classification accuracy. Research on this topic has proposed solutions to be applied in the dissimilarity space \cite{7,10} and/or in the original feature space \cite{4,6}. In this work, we are interested in techniques that fall into the second group. A full review of instance selection methods used in the dissimilarity space can be found in the work by Plasencia-Calaña et al. \cite{9}.

A straightforward instance selection method is the random selection (RS) which seeks $k$ prototypes randomly from the training set without taking into account the class labels. This method can be applied in a stratified fashion, where $k$ prototypes from each class are selected. This allows to produce a more uniform reduced data set with respect to the class distribution.

Other more “intelligent” method seeks to retain points that are closer to the decision boundaries, while removing internal points. One of the earliest methods is the Condensed Nearest Neighbour (CNN) proposed by Hart \cite{20}. This algorithm finds a condensed subset $CS$ from the training set $T$ that correctly classifies every prototype in $T$ using the nearest neighbour (1-NN) rule. This approach starts by randomly selecting one pattern belonging to each class from $T$ and putting them into $CS$. Each remaining sample in $T$ is then classified using the objects in the current $CS$. If a sample in $T$ is misclassified, it is added to $CS$. This process ends when no sample in $T$ is misclassified by $CS$. Nevertheless, this algorithm does not guarantee minimality and both the quality and size of the condensed subset depend on the order in which the training objects are presented to the algorithm.

To overcome the aforementioned issues, Barandela et al. \cite{21} proposed the Modified Selective Subset (MSS) method, which reduces the training set size while preserving the original decision boundaries as much as possible.

## 4 Resampling Techniques

Resampling consists of artificially balancing the original data set, either by over-sampling the minority class and/or by under-sampling the majority class, until the problem classes are approximately equally represented. Both strategies can be applied in any learning system, since they act as a preprocessing phase, allowing the learning system to receive the training objects as if they belonged to a well-balanced data set. Thus, any bias of the system towards the majority class due to the different proportion of examples per class would be expected to be suppressed. The simplest method to increase/reduce of the minority/majority class corresponds to non-heuristic methods that aim to balance the class distribution through the random replication/elimination of positive/negative objects. Nevertheless, these methods have shown important drawbacks. Random over-sampling may increase the likelihood of overfitting, since it makes exact copies of the minority class objects. On the other hand, random under-sampling may discard data potentially important for the classification process. Despite this problem, it has empirically been shown to be one of the most effective resampling methods. In order to overcome these drawbacks, several authors have developed focused resampling algorithms that create balanced data sets in an intelligent way.

Chawla et al. \cite{12} proposed an over-sampling technique that generates new synthetic minority objects by interpolating between several positive examples that lie close
together. This method, called SMOTE (Synthetic Minority Oversampling TTechnique), allows the classifier to build larger decision regions that contain nearby objects from the minority class. From the original SMOTE algorithm, several modifications have been proposed in the literature, most of them pursuing to determine the region in which the positive examples should be generated. For instance, Borderline-SMOTE \[13\] consists of using only positive examples close to the decision boundary, since these are more likely to be misclassified.

Unlike the random method, many proposals are based on a more intelligent selection of the majority class examples to eliminate. For example, Kubat and Matwin \[22\] proposed an under-sampling technique named one-sided selection, that selectively removes only those negative instances that are “redundant” or that “border” the minority class objects (they assume that these bordering cases are noise). In contrast to the one-sided selection technique, the so-called neighborhood cleaning rule emphasizes more data cleaning than data reduction. To this end, Wilson’s editing is used to identify and remove noisy negative objects. Similarly, Barandela et al. \[14\] introduced a method that eliminates not only noisy examples of the majority class by means of Wilsons editing, but also redundant examples through the MSS condensing algorithm.

5 Experimental Setup

Experiments have been carried out over 13 data sets taken from the UCI Machine Learning Database Repository \[23\] and a private library (http://www.vision.uji.es/~sanchez/Databases/). All data sets have been transformed into two-class problems by keeping one original class (the minority class) and joining the objects of the remaining classes (giving the majority class). For example, in Segmentation database the objects of classes 1, 2, 3, 4 and 6 have been joined to shape a unique majority class and the original class 5 has been left as the minority class (see a summary in Table 1).

A stratified five-fold cross validation method has been adopted for the present experiments: each original data set has been randomly divided into five parts or equal (or approximately equal) size. For each fold, four of the parts have been pooled as the training data, and the remaining block has been employed as an independent set. The training sets (in the feature space) have been preprocessed by SMOTE and random under-sampling (RUS) to handle the class imbalance problem. Also, three instance selection methods previously described have been applied over the original training sets (without any preprocessing) and the balanced data sets to gain a reduced representation set: random selection (R), the Condensed Nearest Neighbour (CNN) and the Modified Selective Subset method (MSS). In the case of the random selection method, we have selected 50% (R50) and 100% (R100) of objects from each class. Next, we have computed the dissimilarity matrix $D(T, R)$ by using either imbalanced and balanced training sets with their respective representative sets. Finally, two learners, Fisher and 1-NN classifiers, were applied on the dissimilarity space.

5.1 Performance Evaluation in Class Imbalance Problems

Evaluation of classification performance plays a critical role in the design of a learning system and therefore, the use of an appropriate measure becomes as important as the
Table 1. Data sets used in the experiments

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Positive Examples</th>
<th>Negative Examples</th>
<th>Classes</th>
<th>Majority Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast</td>
<td>81</td>
<td>196</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Ecoli</td>
<td>35</td>
<td>301</td>
<td>8</td>
<td>1,2,3,5,6,7,8</td>
</tr>
<tr>
<td>German</td>
<td>300</td>
<td>700</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Glass</td>
<td>17</td>
<td>197</td>
<td>9</td>
<td>1,2,4,5,6,7,8,9</td>
</tr>
<tr>
<td>Haberman</td>
<td>81</td>
<td>225</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Laryngeal</td>
<td>53</td>
<td>639</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Phoneme</td>
<td>1586</td>
<td>3818</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Pima</td>
<td>268</td>
<td>500</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Scrapie</td>
<td>531</td>
<td>2582</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Segmentation</td>
<td>330</td>
<td>1980</td>
<td>6</td>
<td>1,2,3,4,6</td>
</tr>
<tr>
<td>Spambase</td>
<td>1813</td>
<td>2788</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Vehicle</td>
<td>212</td>
<td>634</td>
<td>4</td>
<td>2,3,4</td>
</tr>
<tr>
<td>Yeast</td>
<td>429</td>
<td>1055</td>
<td>10</td>
<td>1,3,4,5,6,7,8,9,10</td>
</tr>
</tbody>
</table>

Table 2. Confusion matrix for a two-class decision problem

<table>
<thead>
<tr>
<th>Predicted positive</th>
<th>Predicted negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual positive</td>
<td>True Positive (TP)</td>
</tr>
<tr>
<td>Actual negative</td>
<td>False Positive (FP)</td>
</tr>
</tbody>
</table>

The classification accuracy (Acc) evaluates the effectiveness of the learner by its percentage of correct predictions,

\[ Acc = \frac{TP + TN}{TP + FN + TN + FP} \]  

The counterpart of accuracy is the error rate, which evaluates a classifier by its percentage of incorrect predictions.

\[ Err = \frac{FP + FN}{TP + FN + TN + FP} = 1 - Acc \]

Empirical and theoretical evidences show that these measures are strongly biased with respect to data imbalance and proportions of correct and incorrect classifications [24,25,26,27,28]. In a binary decision problem, a learner predicts objects as either positive or negative; if very few examples belong to the positive class, a naive learning system could obtain a very high accuracy by just classifying all objects as negative. However, this is useless in most real domains because the class of interest is generally the positive one. Therefore, evaluators such as accuracy or error rate appear to be inappropriate for class imbalanced data, thus motivating the search for other measures based
Instance Selection and Resampling for Dissimilarity Representation

on some straightforward indexes, which have also been formulated from a $2 \times 2$ confusion matrix as that in Table 2. For example, Kubat and Matwin [22] use the geometric mean of accuracies measured separately on each class,

$$G\text{mean} = \sqrt{TPr \cdot TNr}$$

where $TPr = TP/(TP + FN)$ is the percentage of positive examples that are correctly classified, while, $TNr = TN/(TN + FP)$ is defined as the proportion of negative examples that are correctly classified.

The $G\text{mean}$ is associated to a point on the ROC curve, and the idea is to maximize the accuracies of both classes while keeping them balanced. It can be interpreted as a kind of good trade-off between both rates because a high value occurs when they both are also high, whereas a low value is related to at least one low rate.

5.2 Statistical Significance Tests

A common way to compare two classifiers over a set of problems is the Student’s paired $t$-test. However, this appears to be conceptually inappropriate and statistically unsafe because parametric tests are based on a variety of assumptions (independence, normality and homoscedasticity) that are often violated due to the nature of the problems [29]. In general, the non-parametric tests (e.g., Wilcoxon and Friedman tests) should be preferred over the parametric ones, especially in multi-problem analysis, because they do not assume normal distributions or homogeneity of variance [29,30].

The Friedman test is based on the average ranked performances of a collection of techniques on each data set separately. Under the null-hypothesis, which states that all the algorithms are equivalent, the Friedman statistic can be computed as follows:

$$\chi^2_F = \frac{12N}{K(K+1)} \left[ \sum_j R^2_j - \frac{K(K+1)^2}{4} \right]$$

where $N$ denotes the number of data sets, $K$ is the total number of algorithms and $R_j$ is the average ranks of algorithms. The $\chi^2_F$ is distributed according to the Chi-square distribution with $K - 1$ degrees of freedom, when $N$ (number of data sets) and $K$ (number of algorithms) are big enough. However, it has been demonstrated that the Friedman statistic produces an undesirably conservative effect. In order to overcome the conservativeness, Iman and Davenport [31] proposed a better statistic distributed according to the $F^r$–distribution with $K - 1$ and $(K - 1)(N - 1)$ degrees of freedom,

$$F^r_F = \frac{(N - 1)\chi^2_F}{N(K - 1) - \chi^2_F}$$

When the null-hypothesis is rejected, we can use post-hoc tests in order to find the particular pairwise comparisons that produce statistical significant differences. A post-hoc test compares a control algorithm opposite to the remainder techniques, making possible to define a collection of hypothesis around the control method. The Nemenyi
post-hoc test, which is analogous to the Tukey test for ANOVA, states that the performances of two or more algorithms are significantly different if their average ranks are at least as great as their critical difference (CD) with a certain level of significance:

\[
CD = q_\alpha \sqrt{\frac{K(K+1)}{6N}}
\]

where \( q_\alpha \) is a critical value based on the studentised range statistic divided by \( \sqrt{2} \). For the present set-up, the corresponding critical values are \( q_{0.05} = 3.268 \) and \( q_{0.10} = 3.030 \), for \( \alpha = 0.05 \) and \( \alpha = 0.10 \), respectively.

6 Results and Discussion

Table 3 reports the results, in terms of \( Gmean \), given by Fisher and 1-NN classifiers on the 13 data sets. For each strategy here proposed, the average Friedman is also shown. The technique achieving the best \( Gmean \) on each data set as well as the average Friedman’s ranking is highlighted in bold. From these results, several comments can be drawn:

- In general, when dissimilarity spaces are constructed on balanced datasets (for any instance selection method), the \( Gmean \) values are significantly better than those obtained by using the original training set (without preprocessing).
- The benefits of resampling are much more obvious in the Glass data set, where \( Gmean \) increases from 0.00 (all minority objects were misclassified) to 0.686 with the Fisher classifier.
- Paradoxically, for both 1-NN and Fisher classifiers, the random selection method achieves the best classification results.
- The RUS+R100 strategy has the best Friedman ranking in the case of Fisher, whereas SMOTE with both versions of random selection provide the best average ranking in 1-NN. As claimed by Pekalska and Duin [2], the nearest neighbours classifiers may require a much larger representation set to generate a higher accuracy.
- Although R100 seems a good strategy, it is important to remark that this technique may produce an increase in the computational cost. This problem might grow up if it is combined with an oversampling technique.

In order to check whether there are significantly differences in the results, we computed the Iman-Davenport’s statistic using the Eq. 7 described above. The computation yields \( F_F = 9.585 \) and \( F_F = 4.486 \) for 1-NN and Fisher classifiers, respectively. The critical value for the \( F \) distribution with 12-1=11 and (12-1)(13-1)=132 degrees of freedom considering two levels of confidence, \( \alpha = 0.05 \) and \( \alpha = 0.05 \), are \( F(11,132),0.05 = 1.86 \) and \( F(11,132),0.10 = 1.62 \), so the null hypothesis that all strategies here explored perform equally well can be rejected. Therefore, we can apply Nemenyi’s post hoc test in order to detect the set of strategies that are significantly worse than the control method (the method with the best Friedman’s rank).

The results of the Nemenyi’s post hoc test can be found in Fig. 1. For each classifier and level of confidence, the plot shows the strategies here proposed, which have been
Table 3. Average Gmean results obtained with Fisher and 1-NN classifiers (for each data set, the best case is highlighted in bold)

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Original</th>
<th>SMOTE</th>
<th>RUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast</td>
<td>0.580</td>
<td>0.591</td>
<td>0.578</td>
</tr>
<tr>
<td>German</td>
<td>0.645</td>
<td>0.652</td>
<td>0.634</td>
</tr>
<tr>
<td>Laryngeal2</td>
<td>0.838</td>
<td>0.884</td>
<td>0.818</td>
</tr>
<tr>
<td>Pima</td>
<td>0.669</td>
<td>0.657</td>
<td>0.660</td>
</tr>
<tr>
<td>Scrapie</td>
<td>0.415</td>
<td>0.442</td>
<td>0.422</td>
</tr>
<tr>
<td>Spambase</td>
<td>0.885</td>
<td>0.897</td>
<td>0.897</td>
</tr>
<tr>
<td>Vehicle</td>
<td>0.630</td>
<td>0.626</td>
<td>0.626</td>
</tr>
<tr>
<td>Ecoli</td>
<td>0.773</td>
<td>0.676</td>
<td>0.695</td>
</tr>
<tr>
<td>Glass</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Haberman</td>
<td>0.533</td>
<td>0.534</td>
<td>0.518</td>
</tr>
<tr>
<td>Segmentation</td>
<td>0.907</td>
<td>0.936</td>
<td>0.928</td>
</tr>
<tr>
<td>Yeast</td>
<td>0.667</td>
<td>0.671</td>
<td>0.669</td>
</tr>
<tr>
<td>Phoneme</td>
<td>0.871</td>
<td>0.884</td>
<td>0.883</td>
</tr>
</tbody>
</table>

Average Ranking: 8.62, 7.04, 8.38, 7.35, 3.46, 3.81, 10.46, 6.58, 3.85, 3.04, 7.31, 8.12

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Original</th>
<th>SMOTE</th>
<th>RUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast</td>
<td>0.510</td>
<td>0.533</td>
<td>0.561</td>
</tr>
<tr>
<td>German</td>
<td>0.528</td>
<td>0.527</td>
<td>0.527</td>
</tr>
<tr>
<td>Laryngeal2</td>
<td>0.711</td>
<td>0.738</td>
<td>0.708</td>
</tr>
<tr>
<td>Pima</td>
<td>0.605</td>
<td>0.596</td>
<td>0.597</td>
</tr>
<tr>
<td>Scrapie</td>
<td>0.509</td>
<td>0.510</td>
<td>0.508</td>
</tr>
<tr>
<td>Spambase</td>
<td>0.732</td>
<td>0.734</td>
<td>0.733</td>
</tr>
<tr>
<td>Vehicle</td>
<td>0.561</td>
<td>0.557</td>
<td>0.557</td>
</tr>
<tr>
<td>Ecoli</td>
<td>0.715</td>
<td>0.716</td>
<td>0.708</td>
</tr>
<tr>
<td>Glass</td>
<td>0.541</td>
<td>0.541</td>
<td>0.000</td>
</tr>
<tr>
<td>Haberman</td>
<td>0.571</td>
<td>0.579</td>
<td>0.575</td>
</tr>
<tr>
<td>Segmentation</td>
<td>0.894</td>
<td>0.894</td>
<td>0.891</td>
</tr>
<tr>
<td>Yeast</td>
<td>0.643</td>
<td>0.635</td>
<td>0.638</td>
</tr>
<tr>
<td>Phoneme</td>
<td>0.843</td>
<td>0.844</td>
<td>0.844</td>
</tr>
</tbody>
</table>

Average Ranking: 7.2, 6.5, 7.6, 8.0, 3.6, 3.6, 7.5, 7.8, 4.7, 4.3, 8.1, 9.2

listed in ascending order based in their ranking values (on the y-axis), and the ranking obtained by the Friedman test is displayed on the x-axis. A horizontal dashed line is drawn to represent the end of the best performing technique (the control method). All methods which are on the right side of this line belong to the strategies whose performance is significantly worse than the control method. From these results, in the case of Fisher classifier, the strategies RUS+MSS, CNN, R50 and SMOTE+Cnn perform
In this paper, we have analyzed the effect of the representation set in the dissimilarity space when data are imbalanced. For this purpose, we have evaluated four prototype selection methods and two resampling techniques (one corresponding to under-sampling and one to over-sampling). All these algorithms have also been applied to the data sets before representing them by dissimilarities, with the aim to analyze the influence of having a balanced representation set on the classification performance.

Using the Fisher and 1-NN classifiers, it has been observed that in general, the best classification results in terms of the geometric mean of accuracies are obtained when the training data sets have previously been preprocessed by using some resampling algorithm.

**Acknowledgements.** This work has partially been supported by the Spanish Ministry of Education and Science under grants CSD2007–00018, AYA2008–05965–0596 and TIN2009–14205, the Fundació Caixa Castelló-Bancaixa under grant P1–1B2009–04, and the Generalitat Valenciana under grant PROMETEO/2010/028.
References