Improving minority class prediction using cost-sensitive ensembles

Bartosz Krawczyk, Michal Wozniak, and Gerald Schaefer

Abstract In this paper, we address the problem of dealing with unbalanced datasets in the context of classification, i.e. where some of the classes contain significantly more objects than the other(s). We show that we can this problem by choosing classifiers for a committee of multiple classifier systems. In particular, we propose to design such an ensemble on the basis of a cost of elementary classifiers, given by a cost matrix. To assure the diversity of the ensemble each of the base classifiers is trained on a random subspace. This allows to improve the recognition rate of the minority class, which is typically low when using canonical classifiers. We evaluated our proposed algorithm on a variety of benchmark datasets and show that it significantly outperforms the base cost-sensitive classifier and its boosted version. The results confirm that our approach is a useful tool for dealing with unbalanced datasets.

Key words: cost-sensitive classification, classifier ensemble, multiple classifier system, genetic algorithm.

1 Introduction

The problem of pattern recognition is widespread, having numerous applications. Consequently, many approaches have been introduced in the literature [5] to provide effective and efficient classification systems. However, it is
also well know that according to the no free lunch theory there is no universal method for all decision problems [4].

Among the different approaches, multiple classifier systems (MCSs) have shown much promise [8] as they are able to exploit the strengths of unique elementary classifiers. While there are various important issues when building ensemble classifiers, we will focus on how to create a pool of elementary classifiers and select only the relevant ones from it. For this, combining similar (homogenous) classifiers will typically not lead to a good MCS, but only increase the computational load. Therefore, we are usually looking for a pool of individual classifiers with possibly different components, i.e. heterogenous classifiers. Many algorithms addressing this subject are inspired by researches connected with the methodology of designing reliable software system [10]. Several works introduce different types of diversity measures which, e.g. trying to minimise the possibility of a coincidental failure [12].

A data set is imbalanced if the classification categories are not approximately equally represented. Often real-world data sets are predominately composed of normal examples with only a small percentage of abnormal cases. While the performance of classification algorithms is typically evaluated using predictive accuracy, this is not appropriate when the data is imbalanced.

To overcome the problem of unbalanced datasets, one can for example select fewer instances from the majority class [16] or introduce artificial samples to the dataset [3]. On the other hand, it is also possible to employ a cost-sensitive approach and hence to connect the cost of exploitation with a misclassification penalty.

In this paper, we propose a classifier ensemble design algorithm which is built on the basis of a cost matrix for improved minority class prediction. Our algorithm is based on our earlier work [9], however in contrast to there, costs are associated with a class rather than with individual features.

The rest of this paper is organised as follows. Section 2 outlines the type of multiple classifier system that our approach is based on. Our new algorithm is then introduced in Section 3. Experimental results are reported in Section 4 while Section 5 concludes the paper.

2 Model of pattern recognition task

The aim of pattern recognition is to assign a given object to one of a number of pre-defined categories, on the basis of supplied features of a training set. Although it is important for the performance of a classifier, we do not focus on feature selection in this paper, but assume that the set of features is given by an expert or chosen by a feature selection method [5].

A pattern recognition algorithm $\Psi$ maps the feature space $X$ to the set of class labels $M$.
Improving minority class prediction using cost-sensitive ensembles

\[ \Psi : X \rightarrow M. \]  

(1)

This is established on the basis of examples from a training set or rules given by experts. The training set consists of learning examples, i.e. observations of features together with their correct classifications.

Let’s assume that we have \( n \) classifiers \( \Psi^{(1)}, \Psi^{(2)}, \ldots, \Psi^{(n)} \). For a given object \( x \), each of them decides if it belongs to class \( i \in M = \{1, \ldots, M\} \). The combined classifier \( \bar{\Psi} \) makes a decision based on

\[
\bar{\Psi} \left( \psi^{(1)}(x), \psi^{(2)}(x), \ldots, \psi^{(n)}(x) \right) = \arg \max_{j \in M} \sum_{l=1}^{n} \delta \left( j, \psi^{(l)}(x) \right) w^{(l)} \psi^{(l)}(x),
\]

(2)

where

\[
\delta (j, i) = \begin{cases} 
0 & \text{if } i \neq j \\
1 & \text{if } i = j 
\end{cases}
\]

(3)

and \( w^{(l)} \) is the weight assigned to the \( l \)-th classifier. One of the most used approaches to setting the values of the weights is suggested by Kuncheva [11] and is based on the individual classifier accuracy \( p^{(l)} \). It is defined as

\[
w^{(l)} \propto \log \frac{p^{(l)}}{1 - p^{(l)}}.
\]

(4)

The weights used in Eq. (2) play a key-role in establishing the quality of \( \bar{\Psi} \). Much research has been dedicated to weight configurations, e.g. in [15, 6] the authors proposed to train a fuser. If we set all weights to 1 then Eq. (2) uses the well known majority voting rule. Some alternative way of decision making based on the value of the support function has been suggested in [17]. However, in this paper our objective is to propose a method for composing an ensemble for the decision rule in Eq. 2).

3 Cost-sensitive ensemble

The problem we are addressing is how to select individual classifiers of an ensemble with respect to misclassification cost. Our aim is to create an ensemble with minimal classification error \( P \) within the cost bounds of a cost matrix \( C \). For this purpose, we need to have a pool of base classifiers at our disposal, although we do not focus on the actual classifier training step, since we are interested to appropriately select cost-sensitive classifiers for the ensemble. For this, we propose a method for creating a start-up pool of classifiers for further evaluation using an evolutionary approach.

As a base classifier we have chosen a cost-sensitive classification tree, using the EG2 algorithm [13]. EG2 uses the information cost function (ICF) to select an attribute. For each attribute, the ICF is calculated. The attribute
with the highest ICF value is then used to partition the data at a certain level of the tree. This is based on the misclassification cost rate proposed in [2]. To have a representative pool of classifiers we need to create a set of them. To do this, we use a random subspace approach [7], which randomly divides the feature space into several subspaces and trains individual classifiers in each of them. This ensures that the pool is diverse and contains heterogeneous rather than homogenous classifiers.

For selecting individual classifiers for the ensemble, we employ a genetic algorithm (GA). An individual in the GA population represents a classifier ensemble, and is represented by a binary vector, with 1’s indicating the chosen individual classifiers (i.e., if we have 10 classifiers then '0010110010' would signify that classifiers 3, 5, 6, and 9 are chosen for the ensemble). In addition, we consider the maximum size of the ensemble $V$ (i.e., the number of non-zero weights for fuser) by selecting only individuals with no more than $V$ classifiers for reproduction and fitness evaluation. A weighted voting fusion is used to combine the classifiers, and the ensemble classification error, calculated on the training set, serves as fitness function. Termination conditions can in principle be adjusted, we usually use the number of iterations without result improvement.

Our proposed algorithm is detailed in Algorithm 1.

4 Experimental results

We carried out a series of experiments to evaluate the quality of our proposed method. The main objectives of these experiments were to establish the performance of the proposed method for designing cost-sensitive ensembles, how the size of the chosen ensemble influences the overall quality, and to compare the proposed approach with canonical classifiers.

All experiments were carried out on five different datasets from the UCI repository\(^1\), described in Table 1.

<table>
<thead>
<tr>
<th>dataset</th>
<th>samples</th>
<th>features</th>
<th>majority class [%]</th>
<th>minority class [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hepatitis</td>
<td>78</td>
<td>19</td>
<td>83</td>
<td>17</td>
</tr>
<tr>
<td>Heart Disease</td>
<td>296</td>
<td>13</td>
<td>54</td>
<td>46</td>
</tr>
<tr>
<td>Mushroom</td>
<td>300</td>
<td>22</td>
<td>62</td>
<td>38</td>
</tr>
<tr>
<td>Coil</td>
<td>2500</td>
<td>21</td>
<td>80</td>
<td>20</td>
</tr>
<tr>
<td>MRI</td>
<td>400</td>
<td>15</td>
<td>77</td>
<td>23</td>
</tr>
</tbody>
</table>

\(^1\) http://archive.ics.uci.edu/ml
Algorithm 1 Proposed ensemble classification algorithm.

Input: $V$ (maximum size of the ensemble), $U$ (set of classifiers)
Output: $Q$ (ensemble), $P$ (ensemble classification error)

Create initial population
Select individuals with total size $\leq V$
for all selected individuals do
  Calculate weight according to Eq. (2)
  Evaluate individual fitness functions
  if fitness $< P$ then
    replace $P$
  end if
end for

while termination conditions not satisfied do
  Select pairs for crossover from best-ranked individuals
  Apply crossover operator
  Apply mutation operator
  Select new individuals with total size $\leq V$
  for all selected individuals do
    Calculate weight according to Eq. (2)
    Evaluate new individuals fitness functions
    if fitness $< P$ then
      replace $P$
    end if
  end for
  Create new population
end while

For all experiments we have assumed the following misclassification cost matrix, presented in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>majority class</td>
</tr>
<tr>
<td>true majority class</td>
<td>1</td>
</tr>
<tr>
<td>minority class</td>
<td>50</td>
</tr>
</tbody>
</table>

We ran the experiments with different maximum sizes of an ensemble $V = 1, 3, 5, 7$. A total of 50 EG2 classifiers were trained using the random subspace approach. A normalised average cost of each of the base classifiers was taken as a criterion for ensemble creation.

For the genetic algorithm, the population size was set to 50. We used tournament selection and a two-point crossover operator. Mutation was performed using a simple bit string mutation operator. An elitist selection strategy was utilised for creating new populations where the maximum size of the elite
was set to 10% of the whole population. As termination condition we set 100 iterations without improvement of the final result.

For comparison, we also implemented a single EG2 classifier and a cost-sensitive boosted version of EG2 [14]. All experiments were conducted using standard 10-fold cross validation (10CV).

The results of experiments, described in terms of sensitivity SE, i.e. the fraction of true positives and (true positives+false negatives), and specificity, i.e. the fraction of true negatives and (true negatives+false positives) are given in Table 3 for all datasets and all classifiers.

We used a statistical significance test to compare the results and judge if their differences are statistically significant. For this purpose, we used a combined 5 x 2 cv F Test [1] with a test score the probability of rejecting the null hypothesis that both classifiers have the same error rates. A small difference in error rate implies that the different algorithms construct two similar classifiers with similar error rates; thus the hypothesis should not be rejected. For a large difference, the classifiers have different error rates, and the hypothesis should be rejected. The statistically significant results are bolded in Table 3.

Table 3 Experimental results on the five UCI datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>Hepatitis</th>
<th>Heart</th>
<th>Mushroom</th>
<th>Coil</th>
<th>MRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>EG2</td>
<td>SE 47.35±0.70</td>
<td>68.50±1.45</td>
<td>73.50±1.23</td>
<td>52.05±2.91</td>
<td>60.21±1.46</td>
</tr>
<tr>
<td></td>
<td>SP 68.35±0.95</td>
<td>74.45±2.10</td>
<td>89.10±0.78</td>
<td>89.50±0.65</td>
<td>76.72±2.01</td>
</tr>
<tr>
<td>BoostEG2</td>
<td>SE 55.00±0.70</td>
<td>70.20±0.56</td>
<td>80.00±0.67</td>
<td>57.03±1.00</td>
<td>64.14±0.45</td>
</tr>
<tr>
<td></td>
<td>SP 73.25±0.50</td>
<td><strong>80.80±1.34</strong></td>
<td><strong>92.30±0.20</strong></td>
<td><strong>93.86±0.13</strong></td>
<td><strong>80.04±1.03</strong></td>
</tr>
<tr>
<td>Ensemble V = 1</td>
<td>SE 45.15±1.50</td>
<td>67.40±1.10</td>
<td>79.92±2.11</td>
<td>51.01±2.51</td>
<td>56.41±1.00</td>
</tr>
<tr>
<td></td>
<td>SP 65.50±1.25</td>
<td>69.20±2.90</td>
<td><strong>84.58±1.01</strong></td>
<td>87.01±1.01</td>
<td>76.72±2.65</td>
</tr>
<tr>
<td>Ensemble V = 3</td>
<td>SE 57.25±0.70</td>
<td>69.90±0.74</td>
<td>80.90±0.65</td>
<td>60.04±0.51</td>
<td>62.22±0.91</td>
</tr>
<tr>
<td></td>
<td>SP 71.00±0.50</td>
<td>77.20±1.79</td>
<td>91.58±0.40</td>
<td>92.10±0.76</td>
<td>80.04±0.77</td>
</tr>
<tr>
<td>Ensemble V = 5</td>
<td>SE <strong>63.00±0.45</strong></td>
<td>73.20±0.46</td>
<td><strong>84.05±0.30</strong></td>
<td>64.95±0.23</td>
<td><strong>71.05±0.66</strong></td>
</tr>
<tr>
<td></td>
<td>SP <strong>77.75±0.35</strong></td>
<td><strong>81.20±1.05</strong></td>
<td><strong>93.20±0.14</strong></td>
<td><strong>93.86±0.20</strong></td>
<td>79.20±1.35</td>
</tr>
<tr>
<td>Ensemble V = 7</td>
<td>SE 55.50±0.70</td>
<td><strong>73.20±0.46</strong></td>
<td>80.34±0.75</td>
<td><strong>63.10±0.46</strong></td>
<td><strong>69.04±0.78</strong></td>
</tr>
<tr>
<td></td>
<td>SP 73.10±0.60</td>
<td><strong>81.20±0.80</strong></td>
<td>92.60±0.35</td>
<td><strong>93.856±0.35</strong></td>
<td>77.20±0.60</td>
</tr>
</tbody>
</table>

The results of our experiments clearly demonstrate that our proposed ensemble gives better performance not only compared to a single cost-sensitive classifier but also to its boosted variant. Typical sizes of well performing ensembles vary from 3 to 5 classifiers, whereas bigger committee did not return better results.

5 Conclusions

In this paper, we address the problem of designing an ensemble to generate a multiple classifier system based on a cost matrix, which we employ to
improve minority class recognition. We use a genetic approach to ensemble generation where cost of each individual classifier is defined as its normalised average cost. Our proposed classification system was evaluated on a variety of benchmark datasets and was demonstrated to give significantly better results compared to the underlying single classifier and its boosted variant.

References