Multi-class Imbalanced Data-Sets with Linguistic Fuzzy Rule Based Classification Systems Based on Pairwise Learning

Alberto Fernández[1,⋆], Mara José del Jesus[1], and Francisco Herrera[2]

[1] Dept. of Computer Science, University of Jaén
Tel.:+34-953-212444; Fax:+34-953-212472
{alberto.fernandez,mjjesus}@ujaen.es

[2] Dept. of Computer Science and A.I., University of Granada
herrera@decsai.ugr.es

Abstract. In a classification task, the imbalance class problem is present when the data-set has a very different distribution of examples among their classes. The main handicap of this type of problem is that standard learning algorithms consider a balanced training set and this supposes a bias towards the majority classes.

In order to provide a correct identification of the different classes of the problem, we propose a methodology based on two steps: first we will use the one-vs-one binarization technique for decomposing the original data-set into binary classification problems. Then, whenever each one of these binary subproblems is imbalanced, we will apply an oversampling step, using the SMOTE algorithm, in order to rebalance the data before the pairwise learning process.

For our experimental study we take as basis algorithm a linguistic Fuzzy Rule Based Classification System, and we aim to show not only the improvement in performance achieved with our methodology against the basic approach, but also to show the good synergy of the pairwise learning proposal with the selected oversampling technique.

Keywords: Imbalanced Data-sets, Multi-class Problems, Pairwise Learning, One-vs-One, Oversampling.

1 Introduction

In the research community on imbalanced data-sets [1], recent efforts have been focused on two-class imbalanced problems. However, multi-class imbalanced learning problems appear with high frequency and the correct identification of each kind of concept is equally important for considering different decisions to be taken. In this framework, the solutions proposed for the binary-class problem may not be directly applicable and as a result, there are few works in the specialised literature that cover this issue at present [2].

⋆ Corresponding author.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, LNAI 6178, pp. 89–98, 2010.

Additionally, learning from multiple classes implies a difficulty for Data Mining algorithms, since the boundaries among the classes can be overlapped, which causes a decrease in performance. In this situation, we can proceed by transforming the original multi-class problem into binary subsets, which are easier to discriminate, via a class binarization technique [3,4].

In this contribution we propose a methodology for the classification of multiclass imbalanced data-sets by combining the pairwise learning or one-vs-one (OVO) approach [3] with the preprocessing of instances via oversampling. The idea is to train a different classifier for each possible pair of classes ignoring the examples that do not belong to the related classes, and to apply a preprocessing technique based on oversampling to those training subsets that have a significant imbalance between their classes. Specifically, in order to rebalance the distribution of training examples in both classes, we will make use of the "Synthetic Minority Over-sampling Technique" (SMOTE) [5], which has shown very good results in our previous works on the topic [6,7].

Our objective is to analyse whether this procedure allows a better discrimination of the different classes of the problem, rather than just applying the basic algorithm, and to study the significance of the preprocessing step by contrasting the performance of our methodology against the simple OVO approach. In order to develop this empirical study, we have chosen a linguistic Fuzzy Rule Based Classification System (FRBCSs), the Fuzzy Hybrid Genetics-Based Machine Learning (FH-GBML) algorithm [8]. Furthermore, we have selected 16 multi-class data-sets from the UCI repository [9] and the measure of performance is based on the Probabilistic AUC [10].

This contribution is organised as follows. First, Section 2 presents the problem of imbalanced data-sets, describing its features and the metric we have employed in the context of multiple classes. Next, Section 3 provides a brief introduction to binarization techniques for dealing with multi-class problems, focusing on the pairwise learning approach. In Section 4 we describe the algorithm selected for the study and we present our classification methodology for multi-class imbalanced data-sets based on pairwise learning and oversampling. In Section 5 the experimental framework for the study is established. The experimental study is carried out in Section 6, where we show the goodness of our model. Finally, Section 7 summarises and concludes the work.

2 Imbalanced Data-Sets in Classification

In the classification problem field, the scenario of imbalanced data-sets appears when the numbers of examples that represent the different classes are very different [2]. The minority classes are usually the most important concepts to be learnt, since they represent rare cases or because the data acquisition of these examples is costly. In this work we use the imbalance ratio (IR) [11], defined as the ratio of the number of instances of the majority class and the minority class, to organise the different data-sets according to their IR.

Most learning algorithms aim to obtain a model with a high prediction accuracy and a good generalisation capability. However, this inductive bias towards such a model poses a serious challenge to the classification of imbalanced data. First, if the search process is guided by the standard accuracy rate, it benefits the covering of the majority examples; second, classification rules that predict the positive class are often highly specialised and thus their coverage is very low, hence they are discarded in favour of more general rules, i.e. those that predict the negative class. Furthermore, it is not easy to distinguish between noise examples and minority class examples and they can be completely ignored by the classifier.

Regarding the empirical measure, instead of using accuracy, a more correct metric is considered. This is due to the fact that accuracy can lead to erroneous conclusions, since it doesn't take into account the proportion of examples for each class. Because of this, in this work we use the AUC metric [12], which can be defined as

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2} \tag{1}$$

where TP_{rate} and FP_{rate} are the percentage of correctly and wrongly classified cases belonging to the positive class respectively.

Since this measure has been introduced for binary imbalanced data-sets, we need to extend its definition for multi-class problems. In the specific case of the AUC metric [10], we will compute a single value for each pair of classes, taking one class as positive and the other as negative. Finally we perform the average of the obtained value. The equation for this metric is as follows:

$$PAUC = \frac{1}{C(C-1)} \sum_{j=1}^{C} \sum_{k \neq j}^{C} AUC(j,k) \tag{2}$$

where $AUC(j,k)$ is the AUC (equation (1)) having j as positive class and k as negative class. c also stands for the number of classes. This measure is known as Probabilistic AUC.

3 Reducing Multi-class Problems by Binarization Techniques: One vs. One Approach

Multi-classes imply an additional difficulty for Data Mining algorithms, since the boundaries among the classes can be overlapped, causing a decrease in the performance level. In this situation, we can proceed by transforming the original multi-class problem into binary subsets, which are easier to discriminate, via a class binarization technique [4].

We will make use of the OVO approach [3], which consists of training a classifier for each possible pair of classes ignoring the examples that do not belong to the related classes. At classification time, a query instance is submitted to all binary models, and the predictions of these models are combined into an overall classification [13]. An example of this binarization technique is depicted in Figure 1.

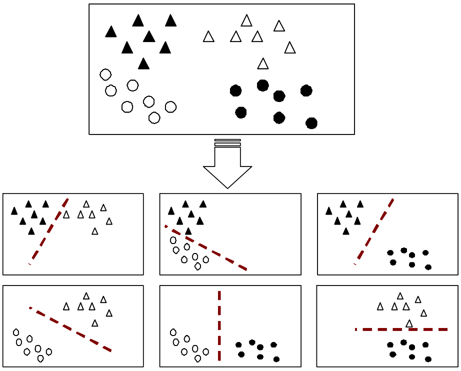

Fig. 1. One-vs-One binarization technique for a 4-class problem

In order to generate the class label, we will use the methodology we have proposed in [14], which considers the classification problem as a decision making problem, defining a fuzzy preference relation with the corresponding outputs of the classifiers. From this fuzzy preference relation, a set of non-dominated alternatives (classes) can be extracted as the solution to the fuzzy decision making problem and thus, the classification output. Specifically, the maximal non-dominated elements of the fuzzy preference relation are calculated by means of the non-dominance criterion proposed by Orlovsky in [15]. In the case of conflict with a given input, i.e. when there are more than one non-dominate value, it remains unclassified due to this ambiguity.

4 Solving Multi-class Imbalanced Data-Sets with Fuzzy Classifiers and Pairwise Learning

In this section we will first describe the FH-GBML algorithm, which will be employed as the base fuzzy model. Then we will present our methodology for dealing with multi-class imbalanced data-sets by means of the combination of multi-classification techniques and preprocessing of instances.

4.1 Fuzzy Hybrid Genetics-Based Machine Learning Rule Generation Algorithm

The FH-GBML method [8] consists of a Pittsburgh approach where each rule set is handled as an individual. It also contains a Genetic Cooperative-Competitive learning (GCCL) approach (an individual represents a unique rule), which is used as a kind of heuristic mutation for partially modifying each rule set.

This method uses standard fuzzy rules with rule weights where each input variable x_i is represented by a linguistic term or label. The system defines 14 possible linguistic terms for each attribute as well as a special “do not care” set.

In the learning process, N_{pop} rule sets are created by randomly selecting N_{rule} training patterns. Then, a fuzzy rule from each of the selected training patterns is generated by probabilistically choosing an antecedent fuzzy set from the 14 candidates ($P(B_k) = \frac{\mu_{B_k}(x_{pi})}{\sum_{j=1}^{14} \mu_{B_j}(x_{pi})}$) and each antecedent fuzzy set of the generated fuzzy rule is replaced with *don't care* using a pre-specified probability.

N_{pop} -1 rule sets are generated by selection, crossover and mutation in the same manner as the Pittsburgh-style algorithm. Next, with a pre-specified probability, a single iteration of the Genetic Cooperative-Competitive-style algorithm is applied to each of the generated rule sets.

Finally, the best rule set is added to the current population in the newly generated (N_{pop} -1) rule sets to form the next population and, if the stopping condition is not satisfied, the genetic process is repeated again.

4.2 Methodology for Dealing with Multi-class Imbalanced Problems with Linguistic Fuzzy Rule Based Classification Systems

Our proposed methodology is defined according to the following two steps:

1. First we will simplify the initial problem into several binary sets, in order to be able to apply those solutions that have been already developed and tested for imbalanced binary-class applications, for example those at data level that change the class size ratio of the two classes via oversampling.

   The advantages of this binarization approach with respect to other techniques, such as confronting one class with the rest ("one-vs-all" [16]), are detailed below:

   - It was shown to be more accurate for rule learning algorithms [17].
   - The computational time required for the learning phase is compensated by the reduction in size for each of the individual problems.
   - The decision boundaries of each binary problem may be considerably simpler than the "one-vs-all" transformation.
   - The selected binarization technique is less biased to obtain imbalanced training-sets which, as we have stated previously in Section 2, may suppose an added difficulty for the identification and discovery of rules covering the positive, and under-represented, samples. Clearly, this last issue is extremely important in our framework.

2. Once we have created all the binary training subsets, we search for those sets that have a significant IR in order to apply the preprocessing step by means of the SMOTE algorithm. According to our previous works on the topic [6], we will consider that the training set is imbalanced if the IR has a value higher than 1.5 (a distribution of 60-40%).

In order to clarify this procedure, the complete process is summarized in Algorithm 1.

Algorithm 1. Procedure for the multi-classifier learning methodology for imbalanced data-sets

```
1. Divide the training set into $C(C-1)/2$ binary subsets for all pairs of classes.
2. For each binary training subset:
   2.1. If IR $> 1.5$
        • Apply SMOTE preprocessing
   2.2. Build a classifier generated with any learning procedure
3. For each input test pattern:
   3.1. Build a fuzzy preference relation $R$ as:
        • For each class $i$, $i = 1, \ldots, m$
            • For each class $j$, $j = 1, \ldots, m$, $j \neq i$
                • The preference degree for $R(i,j)$ is the normalized certainty degree
                  for the classifier associated with classes $i$ and $j$. $R(j,i) = 1 - R(i,j)$
   3.2. Transform $R$ into a fuzzy strict preference relation $R'$.
   3.3. Compute the degree of non-dominance for all classes.
   3.4. The input pattern is assigned to the class with maximum non-dominance value.
```

5 Experimental Framework

In this section we first provide details of the real-world multi-class imbalanced problems chosen for the experimentation and the configuration parameters of the methods, and then we present the statistical tests applied to compare the results obtained with the different approaches.

5.1 Data-Sets and Parameters

Table 1 summarizes the properties of the selected data-sets. It shows, for each data-set, the number of examples (#Ex.), the number of attributes (#Atts.), the number of numerical (#Num.) and nominal (#Nom.) features, the number of classes (#Cl.) and the IR. The *penbased*, *page-blocks* and *thyroid* data-sets have been stratified sampled at 10% in order to reduce their size for training. In the case of missing values (*cleveland* and *dermatology*) we have removed those instances from the data-set. Finally, we must point out that the estimates of the performance were obtained by means of a 5-fold cross validation.

The selected configuration for the FH-GBML approach consists of product T-norm as conjunction operator, together with the Penalised Certainty Factor approach for the rule weight and fuzzy reasoning method of the winning rule. Regarding the specific parameters for the genetic process, we have chosen the following values:

- Number of fuzzy rules: 5 $\cdot d$ rules (max. 50 rules).
- Number of rule sets: 200 rule sets.
- Crossover probability: 0.9.
- Mutation probability: $1/d$.
- Number of replaced rules: All rules except the best-one (Pittsburgh-part, elitist approach), number of rules / 5 (GCCL-part).

Table 1. Summary Description of the Data-Sets

id	Data-set	#Ex.	#Atts.	#Num.	#Nom.	#Cl.	IR
aut	autos	159	25	15	10	6	16.00
bal	balance scale	625	4	4	0	3	5.88
cle	cleveland	297	13	6	7	5	13.42
con	contraceptive method choice	1,473	9	6	3	3	1.89
der	dermatology	366	33	1	32	6	5.55
eco	ecoli	336	7	7	0	8	71.50
gla	glass identification	214	9	9	0	6	8.44
hay	hayes-roth	132	4	4	0	3	1.70
lym	lymphography	148	18	3	15	4	40.50
new	new-thyroid	215	5	5	0	3	4.84
pag	page-blocks	548	10	10	0	5	164.00
pen	pen-based recognition	1,099	16	16	0	10	1.95
shu	shuttle	2,175	9	9	0	5	853.00
thy	thyroid	720	21	6	15	3	36.94
win	wine	178	13	13	0	3	1.5
yea	yeast	1,484	8	8	0	10	23.15

- Total number of generations: 1,000 generations.
- Don't care probability: 0.5.
- Probability of the application of the GCCL iteration: 0.5.

where d stands for the dimensionality of the problem (number of attributes).

For the use of the SMOTE preprocessing technique, we will consider the 5-nearest neighbour to generate the synthetic samples, and balancing both classes to the 50% distribution. In our preliminary experiments we have tried several percentages for the distribution between the classes and we have obtained the best results with a strictly balanced distribution.

5.2 Statistical Tests for Performance Comparison

In this paper, we use the hypothesis testing techniques to provide statistical support for the analysis of the results. Specifically, we will use non-parametric tests, due to the fact that the initial conditions that guarantee the reliability of the parametric tests may not be satisfied, causing the statistical analysis to lose credibility with these type of tests [18,19].

For performing pairwise comparisons between two algorithms, we will apply the Wilcoxon signed-rank test [20]. Furthermore, we consider the average ranking of the algorithms in order to show graphically how good a method is with respect to its partners. This ranking is obtained by assigning a position to each algorithm depending on its performance for each data-set. The algorithm which achieves the best accuracy in a specific data-set will have the first ranking (value 1); then, the algorithm with the second best accuracy is assigned rank 2, and so forth. This task is carried out for all data-sets and finally an average ranking is computed as the mean value of all rankings.

6 Experimental Study

We show the average results in training and test in Table 2, for the three classification schemes analysed in this study, namely the basic approach (Base), the multiclassification approach (OVO) and the multiclassification scheme with oversampling (OVO+SMOTE).

Table 2. Results for the FH-GBML algorithm with the different classification approaches

	Base		OVO		OVO+SMOTE	
Data-set	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}
aut	.7395	.6591	.8757	**.6910**	.8032	.6829
bal	.7178	.7008	.7307	.7109	.7992	**.7296**
cle	.6395	.5577	.7366	**.5664**	.7949	.5584
con	.5852	.5623	.6468	.6201	.6683	**.6294**
der	.7169	.6862	.9746	**.9084**	.9614	.8716
eco	.7564	.7811	.9269	.8201	.9578	**.8321**
gla	.7426	.6920	.8691	.7444	.9375	**.8207**
lym	.8590	.7626	.9349	.8397	.9284	**.8689**
hay	.7979	**.6954**	.9597	.6656	.9663	.6456
new	.9490	.8861	.9967	**.9564**	.9850	.9457
pag	.7317	.6929	.9472	.7862	.9696	**.8552**
pen	.8460	.8340	.9798	**.9508**	.9740	.9387
shu	.7253	.7709	.9319	.8635	.9950	**.9516**
thy	.5198	.4992	.5304	.4993	.9193	**.8763**
win	.9847	.9501	1.000	**.9710**	.9974	.9519
yea	.6456	.6272	.8042	.7438	.8365	**.7442**
Mean	.7473	.7099	.8653	.7711	.9075	**.8064**

We observe that in most cases the best result in test (which is stressed in boldface) corresponds to the one obtained by our OVO+SMOTE methodology. Nevertheless, in order to support the suggestion that our methodology enables an enhancement of the classification ability of the FH-GBML algorithm for imbalanced problems, we will perform a detailed statistical study.

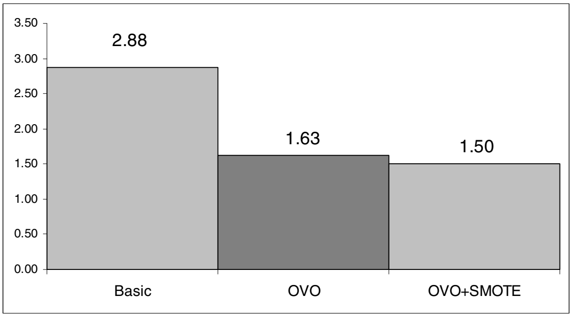

Fig. 2. Average ranking for the FH-GBML method with the different classification schemes

Table 3. Wilcoxon signed-ranks test. R^+ corresponds to the sum of the ranks for the OVO+SMOTE method and R^- to the Basic and OVO classification schemes.

Comparison	R^+	R^-	p-value	Hypothesis ($\alpha = 0.05$)
OVO+SMOTE vs. Basic	131.0	5.0	0.001	Rejected for OVO+SMOTE
OVO+SMOTE vs. OVO	88.0	48.0	0.301	Not Rejected

First, Figure 2 shows the average ranking computed for the different classification schemes, where we can observe that OVO+SMOTE is the best option, whereas the basic FH-GBML approach obtains the worst ranking with a much higher value than the former.

Next, we perform a Wilcoxon test (Table 3) to contrast the different approaches that are being studied. The first conclusion extracted from the result of this test is that our methodology is actually better suited for imbalanced data-set with multiple classes than the basic learning algorithm. Also, we observe that the application of the oversampling step enables the obtention of better results than applying the binarization scheme directly over the original training data, as suggested by both the higher sum of the ranks in favour of our methodology and the average results in Table 2.

The study carried out allow us to discuss several issues as future work:

1. The inclusion of different Machine Learning algorithms to analyse the robustness of our methodology.
2. A comparative study of several preprocessing techniques (oversampling, undersampling and hybrid approaches).
3. A detailed study regarding the IR of the algorithms and the goodness of the application of preprocessing in each case and the definition of a precise threshold in order to rebalance the binary training data.

7 Concluding Remarks

In this paper we have presented a new methodology for the classification of multi-class imbalanced data-sets using a combination of pairwise learning and preprocessing of instances. This methodology divides the original problem into binary-class subsets which are rebalanced using the SMOTE algorithm when the IR between the corresponding classes is higher than a threshold.

We have tested the quality of this approach using the FH-GBML algorithm, a linguistic FRBCSs, for which the experimental results support the goodness of our methodology as it generally outperforms the basic and pairwise learning multi-classifier approach.

Acknowledgment

This work had been supported by the Spanish Ministry of Science and Technology under Projects TIN2008-06681-C06-01 and TIN2008-06681-C06-02, and the Andalusian Research Plan TIC-3928.

References

1. Chawla, N.V., Japkowicz, N., Kolcz, A.: Editorial: special issue on learning from imbalanced data sets. SIGKDD Explorations 6(1), 1–6 (2004)
2. Sun, Y., Wong, A.K.C., Kamel, M.S.: Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence 23(4), 687–719 (2009)
3. Hastie, T., Tibshirani, R.: Classification by pairwise coupling. The Annals of Statistics 26(2), 451–471 (1998)
4. Allwein, E.L., Schapire, R.E., Singer, Y.: Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of Machine Learning Research 1, 113–141 (2000)
5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over–sampling technique. Journal of Artificial Intelligent Research 16, 321–357 (2002)
6. Fernández, A., García, S., del Jesus, M.J., Herrera, F.: A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data–sets. Fuzzy Sets and Systems 159(18), 2378–2398 (2008)
7. Fernández, A., del Jesus, M.J., Herrera, F.: On the 2-tuples based genetic tuning performance for fuzzy rule based classification systems in imbalanced data-sets. Information Sciences 180(8), 1268–1291 (2010)
8. Ishibuchi, H., Yamamoto, T., Nakashima, T.: Hybridization of fuzzy GBML approaches for pattern classification problems. IEEE Transactions on System, Man and Cybernetics B 35(2), 359–365 (2005)
9. Asuncion, A., Newman, D.: UCI machine learning repository. University of California, Berkeley (2007), http://www.ics.uci.edu/~mlearn/MLRepository.html
10. Hand, D.J., Till, R.J.: A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning 45(2), 171–186 (2001)
11. Orriols-Puig, A., Bernadó-Mansilla, E.: Evolutionary rule–based systems for imbalanced datasets. Soft Computing 13(3), 213–225 (2009)
12. Huang, J., Ling, C.X.: Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on Knowledge and Data Engineering 17(3), 299–310 (2005)
13. Hüllermeier, E., Brinker, K.: Learning valued preference structures for solving classification problems. Fuzzy Sets and Systems 159(18), 2337–2352 (2008)
14. Fernández, A., Calderón, M., Barrenechea, E., Bustince, H., Herrera, F.: Enhancing fuzzy rule based systems in multi-classification using pairwise coupling with preference relations. In: EUROFUSE '09 Workshop on Preference Modelling and Decision Analysis (EUROFUSE '09), pp. 39–46 (2009)
15. Orlovsky, S.A.: Decision-making with a fuzzy preference relation. Fuzzy Sets and Systems 1, 155–167 (1978)
16. Rifkin, R., Klautau, A.: In defense of one-vs-all classification. Journal of Machine Learning Research 5, 101–141 (2004)
17. Fürnkranz, J.: Round robin classification. Journal of Machine Learning Research 2, 721–747 (2002)
18. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30 (2006)
19. García, S., Herrera, F.: An extension on "Statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. Journal of Machine Learning Research 9, 2677–2694 (2008)
20. Sheskin, D.: Handbook of parametric and nonparametric statistical procedures, 2nd edn. Chapman & Hall/CRC, Boca Raton (2006)