A Hybrid Approach Handling Imbalanced Datasets

Paolo Soda

University Campus Bio-Medico of Rome, Integrated Research Centre, Medical Informatics & Computer Science Laboratory, Rome, Italy
p.soda@unicampus.it

Abstract. Several binary classification problems exhibit imbalance in class distribution, influencing system learning. Indeed, traditional machine learning algorithms are biased towards the majority class, thus producing poor predictive accuracy over the minority one. To overcome this limitation, many approaches have been proposed up to now to build artificially balanced training sets. Further to their specific drawbacks, they achieve more balanced accuracies on each class harming the global accuracy. This paper first reviews the more recent method coping with imbalanced datasets and then proposes a strategy overcoming the main drawbacks of existing approaches. It is based on an ensemble of classifiers trained on balanced subsets of the original imbalanced training set working in conjunction with the classifier trained on the original imbalanced dataset. The performance of the method has been estimated on six public datasets, proving its effectiveness also in comparison with other approaches. It also gives the chance to modify the system behaviour according to the operating scenario.

1 Introduction

Class imbalance is considered a crucial issue in machine learning and data mining since most learning systems cannot cope with the large difference between the number of instances belonging to each class. A training set (TS) is imbalanced when one of the classes is largely under-represented in comparison to the others. According to previous works, we consider here only binary problems, namely positive or negative. The former belong to the minority class, whereas the latter to the majority one. Traditional algorithms are biased towards the majority class, resulting in poor predictive accuracy over the minority one. Since classifiers are designed to minimise errors over training samples, they may ignore classes composed of few instances. The relevance of learning with imbalanced TS is emphasised observing that it exists in a large number of real-world domains, such as text classification, medical diagnosis, fraud detection, oil spills in satellite images of the sea surface.

Research efforts dealing with imbalanced TSs in supervised learning can be traced back to the following four categories: (i) undersampling the majority class so as to match the size of the other class; (ii) oversampling the minority class

P. Foggia, C. Sansone, and M. Vento (Eds.): ICIAP 2009, LNCS 5716, pp. 209–218, 2009.

so as to match the size of the other class; (iii) internally biasing the discriminating process so as to compensate for the class imbalance; (iv) multi-experts system (MES) composed of multiple balanced classifiers. Despite several existing proposals [1,2,3,4,5,6,7,8,9,10,11,12], some issues remained opened. For instance, balancing the recognition accuracies achieved for each class very often decreases the global accuracy.

In order to overcome present limitations, we propose here an hybrid approach that uses the reliability of each classification act to combine the original imbalanced classifier (IC) with a MES trained on balanced subsets of the TS. The approach has been successfully tested on several public datasets, showing that it can overcome the drawbacks of the existing methods. A second contribution of the paper consists in reviewing the literature, particularly focusing on the MES approach, which has been proposed quite recently in the field of imbalanced datasets.

2 Background

This section first discusses performance metrics, then reviews the literature and finally presents the notion of classification reliability.

Performance measures. The confusion matrix is used to evaluate the performance of classification systems since its element describe the behaviour of the system. With reference to Table 1, which shows the confusion matrix for a two-classes problem, we denote as $n^- = FP + TN$ and $n^+ = TP + FN$ the number of samples in the negative and positive classes, respectively.

Classification accuracy is the traditional performance measure of a pattern recognition system. It is defined as $acc = (TP + TN)/(n^- + n^+)$. However, in case of imbalanced application the performance of a learning algorithm can not be measured in terms of classification accuracy only. For instance, consider a domain where the 3% of samples are positive: in this case, labelling all test patterns as negative will result in an accuracy of 97%, i.e. failing on all positive cases which is clearly meaningless.

Table 1. Confusion matrix for a two-classes problem

	Actual positive	Actual negative
Hypothesise positive	True Positive (TP)	False Positive (FP)
Hypothesise negative	False Negative (FN)	True Negative (TN)

Indeed, when the prior class probabilities are very different, measuring only the accuracy may lead to misleading conclusions since it is strongly biased in favor of the majority class. Such an observation can be easily explained observing that class distribution is the relationship between the first and the second column of the confusion matrix. Any performance measure based on values from both columns will be inherently sensitive to class skews, as accuracy is.

To this end, the most common performance metric for imbalanced datasets is the geometric mean (g) of accuracies, defined as [3]:

$$g = \sqrt{acc^{+} \cdot acc^{-}} \tag{1}$$

where $acc^{+} = \frac{TP}{TP+FN}$ is the *True Positive Rate* and $acc^{-} = \frac{TN}{TN+FP}$ is the*True Negative Rate*. The idea of such an estimator is to maximise the accuracy on each class while keeping these accuracies balanced. For instance, a high value of acc^{+} by a low acc^{-} will result in a low g. Notice that such a measure is non-linear: a change in one of the two parameters has a different effect on g depending upon its magnitude. As an example, the smaller the value of acc^{+}, the greater the variation of g.

Techniques for Handling Imbalanced Datasets. This section reviews the four existing methods coping with imbalanced datasets, which have been introduced in the introduction.

Undersampling the majority class and oversampling the minority one so as to match the size of the other class have received great attention [1,2,3,4,6]. These methods resize the TS making the class distribution more balanced. Nevertheless, both have shown relevant drawbacks. On the one hand, undersampling the TS may remove potentially useful data. On the other hand, oversampling increases the number of samples of the minority class through the random replication of its elements, thus increasing the likelihood of overfitting [2,3]. Furthermore, it increases the time required to train the classifier since TS size grows up. Although this paper does not aim at providing a deep review of existing approaches of such types, it is worth observing that several papers have proposed various methods to reduce the limitations of both under and over sampling methods, such as [2,3,13,14].

Another approach to manage skewed TS consists in internally biasing the discrimination-based process so as to compensate for class imbalance. In [7] the authors proposes a weighted distance function to be used in the classification phase that compensates the TS imbalance without altering the class distribution since the weights are assigned to the respective classes and not to the individual examples. In [8] the authors assigned different weights to prototypes of the different classes. Ezawa et al. [15] biased the classifier in favour of certain attribute relationship, whereas Eavis et al. [9] presented a modified auto-encoder that allows for the incorporation of a recognition component into the conventional Multi-Layer Perceptrons (MLP) mechanism.

The last approach coping with imbalanced TS is based on multi-experts system, also known as ensemble of classifiers, where each composing classifier is trained on a subset of the majority class and on the whole minority class [10,11,12]. The idea is based on the widely accepted result that a MES approach generally produces better results than those obtained by individual composing experts, since different features and recognition systems complement each other in classification performance. Indeed, the MES takes advantage of the strengths of the single experts, without being affected by their weaknesses [16,17]. Furthermore, constructing balanced subsets of the original TS avoids the drawbacks of under and oversampling.

In [10] the authors generated as many training subsets as required to get balanced subsets from the given TS. The number of subsets was determined by the difference between the number of samples in majority and minority classes. Each classifier is trained with a learning set consisting of all samples of the minority class and the same number of training instances randomly selected from those of the majority one. As base classifiers, they employed Nearest Neighbour (NN) classifiers, combining their outputs by majority voting (MV). Their proposal has been tested on four datasets (Phoneme, Satimage, Glass and Vehicle) taken from the UCI Repository [18]. With respect to the original imbalanced classifier (IC), system performance, expressed in terms of geometric mean of accuracies, improves for three of the four datasets. Furthermore, in two cases the performance improves also in comparison with the approach reported in [3].

Molinara et al. [11] presented an approach similar to the previous one. In order to build a classifiers ensemble, they tested two splitting strategies which are based on clustering and random selection of the samples of the majority class. As base classifiers they have adopted Gentle AdaBoost, whereas dynamic selection, mean and MV have been used as criteria to aggregate individual decisions. They applied the approach to detect microcalcifications on a dataset of digital mammograms publicly available. The results reported in terms of accuracies on positive and negative samples showed that the former increases with respect to IC, whereas the latter decreases. Furthermore, such a paper studied how the behaviour of the MES system varies with respect to the number of base classifiers, in the range between one (i.e., the original imbalanced classifier) up to the ratio between majority and minority class cardinalities. Best performance has been achieved when the MES is composed of a number of base classifiers equal to the ratio between the number of majority and minority class samples.

In [12] the authors proposed a MES composed by a fixed number of base classifiers independently of sample distribution in the TS. As base learning systems they employed 5-NN, C4.5 decision tree and Naïve Bayes combined via the MV rule. The approach has been tested on eight UCI datasets showing that the MES improves geometric mean and accuracy on positive samples with respect to IC, while it decreases accuracy on negative patterns. Furthermore, it compares favorably with regard to other methods for handling imbalanced datasets, such as those based on the variation of weights assigned to the classes [19].

From the review reported so far, it is apparent that MES is a suitable alternative to cope with imbalanced dataset, outperforming also other approaches. However, most of the available methods including the MES approach, improve the accuracy on positive samples to the detriment of accuracy on negative samples, increasing g and lowering the global accuracy. Although the latter is not the unique measure to compare different methods handling imbalanced TS, nevertheless it is the ultimate parameter representing the global performance of a classifier. Moreover, to our knowledge, all the approaches provide “fixed” performance, i.e. certain values of g and acc, which are independent from the working scenario.

In this respect, this paper proposes a method that can overcome this limitations. Indeed, not only it can increase both g and acc in comparison with IC and MES approaches, but it also permits to tune the working point of the system on the basis of the operating context.

Reliability Estimation. Exploiting information derived from classifier output permits to properly estimates the reliability of the decision of each classification act [16,17]. Reliability takes into account the many issues that influence the achievement of a correct classification, such as the noise affecting the samples domain or the difference between the objects to be recognized and those used to train the classifier. The reliability assigned to a sample x, denoted as $\phi(x)$ in the following, typically varies in $[0,1]$. A low value of $\phi(x)$, i.e. $\phi(x) \rightarrow 0$, means that the crisp label allotted to x should be wrong. On the other side, $\phi(x) \rightarrow 1$ implies that the recognition system is more likely to provide a correct classification.

3 Methods

The original imbalanced classifier tends to classify test instances as negative since it is trained with a TS skewed toward that class. In this case, misclassifications mostly affect instances belonging to the minority class. Hence, it is reasonable to assume that IC classifier labels minority class samples with low reliability. However, this criterion is not sufficient to label a sample as positive, since low reliability can be measured also for majority class instances.

Previous works demonstrated that a MES trained on balanced subsets of the original TS better recognises minority class patterns than IC [10,11,12], as discussed in section 2. Based on these observations, the proposed method works as follows. When the reliability of the decision taken by IC on sample x is low, x is classified using the label provided by the MES, otherwise x is assigned to the class given by IC. To formally present this criterion let us denote as $O_{IC}(x)$, $O_{MES}(x)$ and $O(x)$ the labels provided by IC, MES and the overall system, respectively, and let t be a threshold value ranging in $[0,1]$. On this basis, the final label is given by:

$$O(x) = \begin{cases} O_{IC}(x) & \text{if } \phi(x) \geq t \\ O_{MES}(x) & \text{otherwise} \end{cases} \tag{2}$$

Since using either IC or MES depends on the reliability of sample classification, to refer to such a method we will use the term *Reliability-based Balancing* (RbB) in the rest of the paper. Fig. 1 graphically describes the RbB method, where the RbB block selects one of its inputs comparing $\phi(x)$ with the threshold t.

To measure the performance of the RbB system, we estimate both g and acc since the former represents the balancing between the two classes accuracies, while the latter the global performance of the classification system, respectively. In a plot where g and acc correspond to the X and Y axes, varying the threshold t in the interval $[0,1]$ returns a set of points that can be used to generate a

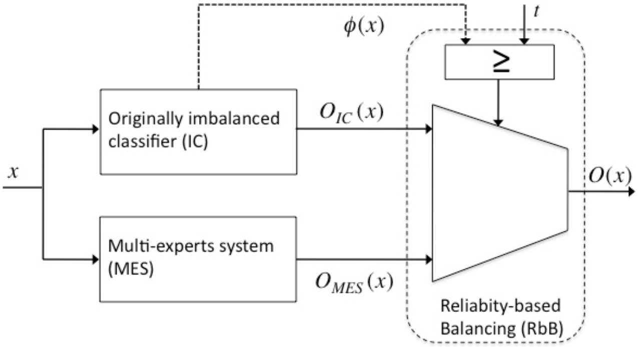
$\phi(x)$
t
≥
Originally imbalanced classifier (IC)
$O_{IC}(x)$
x
$O(x)$
Multi-experts system (MES)
$O_{MES}(x)$
Reliabity-based Balancing (RbB)

Fig. 1. Schematic representation of the proposed method handling with imbalanced TS. The sample x is given to IC, which provides both the label $O_{IC}(x)$ and the reliability measure $\phi(x)$.

curve. The curve extrema are given by the points corresponding to IC and MES performance, i.e. $t = 0$ and $t = 1$, respectively. Fig. 2 reports an instance of such a curve, whose data come from two experiments described in section 4. It is straightforward noting that the ideal point is $(1, 1)$; informally, the nearer the curve to this point, the better the performance obtained.

4 Experimental Evaluation

In this section, we experimentally prove that the RbB method can be used profitably with imbalanced dataset. To this aim, we perform tests on six UCI datasets [18], belonging to real-world problems. They show a variability with respect to the number of features and samples as well as in class distribution (Table 2). According to a general practice reported in previous works on imbalanced TS, for datasets having originally more than two classes we choose the class with fewer instances as the minority one (i.e. positive) and collapsed the others into the negative class. Hence, in the Glass set the problem was transformed to discriminate class 7 against all other classes, and in the Ecoli dataset the task consists in classifying class 4 against the others.

Support Vector Machine (SVM) with a Radial Basis Function (RBF) Kernel has been used as classifier for both IC and MES configuration. To evaluate the reliability of its decisions we use the distance of the pattern x from the optimal separating hyperplane in the feature space induced by the chosen kernel [20].

In order to build the MES composed of individual balanced classifiers, we randomly generate as many training subsets as is the ratio between the cardinalities of minority and majority classes, since this choice has been successfully adopted in previous works [10,11]. Hence, each base classifier is trained with a learning set composed of all negative samples and an approximately equal number of

Table 2. Summary of the used datasets

Dataset	Number of Samples	Number of features	Class Distribution (%) (minority, majority)
Ecoli	336	7	(10.4, 89.6)
Glass	214	9	(7.9, 92.1)
Hepatitis	155	19	(20.8, 79.2)
Pima	768	8	(34.8, 65.2)
Phoneme	5404	5	(29.4, 70.6)
Breast Cancer Wisconsin	699	9	(34.5, 65.5)

positive instances. The outputs of such classifiers are combined by the Weighted Voting (WV) rule, which weights the classification of each expert about the class by a reliability parameter.

4.1 Results and Discussion

Tests were performed using 5-folds cross validation and averaging the results over the runs. As performance parameters we measure the global accuracy, *acc*, and the geometric mean of accuracies, g, as motivated in section 3.

Table 3 reports for each dataset the results achieved by the originally imbalanced classifier (IC) and by the MES composed of several balanced classifiers.

Table 3. Results of the original imbalanced classifier (IC) and of the MES composed of several balanced classifiers as well as of RbB method, reported in terms of accuracy (*acc*) and geometric mean (g)

Dataset	Classifier	Accuracy (*acc*)	Geometric Mean (g)
Ecoli	IC	91.9%	72.4%
	MES Random	82.9%	88.9%
	RbB	92.5%	91.1%
Glass	IC	94.8%	87.4%
	MES Random	90.6%	87.6%
	RbB	94.8%	89.9%
Hepatitis	IC	85.6%	76.5%
	MES Random	76.6%	78.4%
	RbB	86.2%	78.8%
Pima	IC	77.7%	71.5%
	MES Random	74.3%	74.6%
	RbB	76.0%	74.8%
Phoneme	IC	75.6%	64.8%
	MES Random	69.9%	73.5%
	RbB	78.2%	75.1%
Breast Cancer Wisconsin	IC	96.5%	95.9%
	MES Random	97.0%	97.0%
	RbB	96.7%	97.1%

These results confirm previous findings since in five out of the six considered datasets balanced MES outperforms IC in terms of g, while the global accuracy decreases.

The same table reports the performance achieved applying the proposed method. The reported values correspond to the point of the curve closest to the upper right corner of the curve, which represents the best system performance as observed in section 3.

Figures 2 shows in the left and right panel the curves achieved varying t in case of Phoneme and Pima datasets , respectively. Other plots, where we observed similar trend, have been omitted for space reason. Points marked by IC, MES represent the performance of IC and MES solution, whereas point labelled as A corresponds to the performance of the RbB method reported in Table 3. In the left panel it is straightforward finding out the point corresponding to the best performance, which is the closest to the upper right corner. Right panel shows a less favourable case, where RbB does not outperform both IC and MES solutions, since its global accuracy is slightly smaller than IC one, but it provides a value of b higher than other solutions.

Data experimentally prove that RbB criterion has the valuable capability of improving both g and acc with respect to IC and MES in four out of the six datasets. In case of Breast Cancer dataset the MES itself outperforms IC and therefore RbB does not introduce any improvements. Furthermore, for Pima dataset RbB rule improves g in comparison with both IC and MES, while acc increases only with respect to MES configuration. Anyway, we deem that such findings can be regarded as positive results.

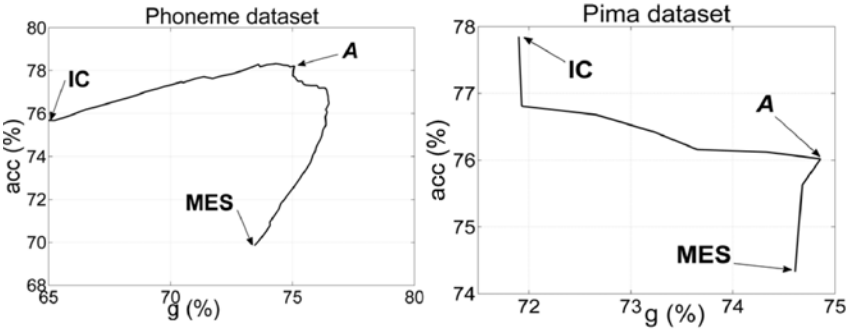

Fig. 2. Diagram of the RbB performance with respect to threshold variation in case of Phoneme and Pima datasets (left and right panel, respectively)

5 Conclusions

Methods handling imbalanced TS increase the geometric mean of accuracies harming the global accuracy. To overcome such a drawback, we have presented an approach integrating IC reliability and label with the decision taken by a

MES. This method increases in most cases both performance metrics as result of the estimation of the classification reliability. Furthermore, it returns a curve that can be used to set the operating point of the classification system on the basis of the domain peculiarities.

The interesting results achieved motivate us to further investigate such an approach, to develop an optimal rule to set the operating point, to analyse classifiers aggregation rules in case of imbalanced TS, and to extend the approach to problems with more than two classes. Furthermore, we are planning to compute the optimal threshold value on a validation set to provide a more careful evaluation of method performance.

References

1. Batista, G.E., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter 6(1), 20–29 (2004)
2. Chawla, N.V., Bowyer, K.W., et al.: Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16(3), 321–357 (2002)
3. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: One-sided selection. In: Machine Learning-International Workshop Then Conference, pp. 179–186. Morgan Kaufmann Publishers, Inc., San Francisco (1997)
4. Jo, T., Japkowicz, N.: Class imbalances versus small disjuncts. ACM SIGKDD Explorations Newsletter 6(1), 40–49 (2004)
5. Ling, C.X., Li, C.: Data mining for direct marketing: Problems and solutions. In: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, pp. 73–79 (1998)
6. Weiss, G.M., Provost, F.: Learning when training data are costly: the effect of class distribution on tree induction. Journal of Artificial Intelligence Research 19, 315–354 (2003)
7. Barandela, R., Sanchez, J.S., Garca, V., Rangel, E.: Strategies for learning in class imbalance problems. Pattern Recognition 36(3), 849–851 (2003)
8. Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., Brunk, C.: Reducing misclassification costs. In: Proceedings of the Eleventh International Conference on Machine Learning, pp. 217–225 (1994)
9. Eavis, T., Japkowicz, N.: A recognition-based alternative to discrimination-based multi-layer perceptrons. In: AI 2000: Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence, pp. 280–292 (2000)
10. Barandela, R., Valdovinos, R.M., Sánchez, J.S.: New applications of ensembles of classifiers. Pattern Analysis & Applications 6(3), 245–256 (2003)
11. Molinara, M., Ricamato, M.T., Tortorella, F.: Facing imbalanced classes through aggregation of classifiers. In: ICIAP 2007: Proceedings of the 14th International Conference on Image Analysis and Processing, pp. 43–48 (2007)
12. Kotsiantis, S., Pintelas, P.: Mixture of expert agents for handling imbalanced data sets. Annals of Mathematics, Computing and Teleinformatics 1(1), 46–55 (2003)
13. Japkowicz, N.: Concept-learning in the presence of between-class and within-class imbalances. In: AI 2001: Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence, pp. 67–77 (2001)

14. Laurikkala, J.: Improving Identification of Difficult Small Classes by Balancing Class Distribution. Springer, Heidelberg (2001)
15. Ezawa, K., Singh, M., Norton, S.: Learning goal oriented bayesian networks for telecommunications risk management. In: Machine Learning-International Workshop Then Conference, pp. 139–147 (1996)
16. Cordella, L.P., Foggia, P., Sansone, C., Tortorella, F., Vento, M.: Reliability parameters to improve combination strategies in multi-expert systems. Pattern Analysis & Applications 2(3), 205–214 (1999)
17. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(3), 226–239 (1998)
18. Asuncion, A., Newman, D.J.: UCI machine learning repository (2007)
19. Domingos, P.: Metacost: a general method for making classifiers cost-sensitive. In: KDD 1999: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 155–164. ACM, New York (2000)
20. Fumera, G., Roli, F.: Support Vector Machines with Embedded Reject Option. LNCS, pp. 68–82 (2002)