ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

A post-processing strategy for SVM learning from unbalanced data

Haydemar Núñez[1], Luis Gonzalez-Abril[2], and Cecilio Angulo[3] *

1- Artificial Intelligence Laboratory - Central University of Venezuela
Faculty of Sciences. Los Ilustres, Urb. Valle Abajo. Caracas 1020-A - Venezuela

2- Qualitative Methods Research Group - University of Seville
Avda Ramon y Cajal s/n, 41018 Seville - Spain

3- Knowledge Engineering Research Group - Technical University of Catalonia
Avda Víctor Balaguer 1, 08800 Vilanova i la Geltrú - Spain

Abstract. Standard learning algorithms may perform poorly when learning from unbalanced datasets. Based on the Fisher's discriminant analysis, a post-processing strategy is introduced to deal datasets with significant imbalance in the data distribution. A new bias is defined, which reduces skew towards the minority class. Empirical results from experiments for a learned SVM model on twelve UCI datasets indicates that the proposed solution improves the original SVM, and they also improve those reported when using a z-SVM, in terms of g-mean and sensitivity.

1 Introduction

In bi-classification, learning from unbalanced datasets occurs when the provided datasets to the learning algorithms contain many examples for a class, but very few for the other. Hence, a good model is difficult to generate using traditional classification techniques since the objective functions used for learning the classifiers typically tend to favor the larger, usually less important, class [1]. This situation arises in domains such as medical diagnosis, text classification, credit card fraud detection, intrusion in communication networks, and others.

Support Vector Machine [2] is an attractive option for dealing with unbalanced datasets because its learning mechanism usually considers a small subset of patterns to build the classification model. However, like other learning machines that build these models, SVM aims to minimize the error on the entire dataset, so it is inherently biased towards the majority class. Thus, SVM will learn to classify all examples as belonging to this class for a severe imbalance.

This paper is focused on SVM learning from unbalanced datasets. In particular, SVM performance will be improved by introducing a new bias for the induced classification function. Previously, some general strategies and metrics to evaluate classifiers are introduced in the next section. Section 3 presents the problem for the case of SVM learning. A new SVM bias is defined in Section 4, as a novel post-processing strategy. Experimentation and results are presented in Section 5. Finally, some conclusions and future work are provided.

*This research has been partly supported by the ARTEMISA project, TIN2009-14378-C02-01, from the Spanish Ministry of Science and Technology. Cecilio Angulo acknowledges the I3 grant from the General Program for Research Intensification, by Universitat Politécnica de Catalunya, 2008-2011.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

2 Binary learning from unbalanced datasets

In binary learning from unbalanced datasets, the class with fewer examples is known as the minority class or positive, while the other class is called the majority class or negative. Reasons leading to the imbalance between classes are the nature of the problem or the cost in obtaining data. A particular case is when binary classifiers, like SVM, are used to solve multi-class classification problems by considering the standard one versus rest strategy.

With unbalanced datasets, often the simplest learned hypothesis is to classify all examples as negative. To overcome this problem different strategies have been proposed [1, 3], including: sampling methods; generating artificial data, either by over-sampling the minority class, or under-sampling the majority class; cost-sensitive learning, over-weighting errors on the minority class; ensemble methods, trained from learning sets with different data distributions; post-processing by tuning the learned classification function to improve performance on minority class; modified traditional algorithms and new algorithms.

Furthermore, the usual classification error and predictive accuracy metrics are not appropriated when the prior probabilities of the classes are very different because they do not consider costs from wrong classifications and thus they are very sensitive to the bias between classes [1]. Therefore, other measures of assessment based on the confusion matrix are considered: $Sensitivity = \frac{tp}{tp+fp}$ and $Recall = \frac{tp}{tp+fn}$, with tp being true positives, fp false positives and fn false negatives. Sensitivity is a measure of accuracy among examples classified as positive and Recall is a measure of completeness. Both measures, in contrast to Error and Accuracy, are not sensitive to changes in data distribution and can effectively evaluate the classification performance in unbalanced learning scenarios. Hence, the g-mean measure, defined as,

$$g - mean = \sqrt{\frac{tp}{tp+fn} \cdot \frac{tn}{tn+fp}} \tag{1}$$

can measure the accuracy of both classes with a good trade-off. Other evaluation techniques are the ROC curve analysis (Receiver Operating Characteristics) and the analysis based on Sensitivity and Recall curves.

3 SVM learning from unbalanced datasets

SVMs are learning machines which implement the structural risk-minimization inductive principle to obtain good generalization on a limited number of learning patterns [2]. This theory was developed on the basis of a separable binary classification problem where the optimization criterion is the width of the margin between the positive and negative examples. The extension of binary classification to multi-classification is currently an on-going research issue, however the binary ad-hoc methods of one-versus-rest SVMs to solve the multi-class problem still prevails due to in general good performance and manageable optimization.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Let $\mathcal{Z} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a training set, with $x_i \in \mathcal{X}$ as the input space, $y_i \in \mathcal{Y} = \{\theta_1, \theta_2\} = \{+1, -1\}$ the output space, and $z_i = (x_i, y_i)$. Let $\phi : \mathcal{X} \to \mathcal{F}$, $\mathrm{x} = \phi(x)$, be a feature mapping with a dot product denoted by $\langle \cdot, \cdot \rangle$. A binary linear classifier, $f(x) = \langle \mathrm{x}, \mathrm{w} \rangle + b$, is sought with $\mathrm{w} \in \mathcal{F}$, $b \in \mathbb{R}$, and where outputs are obtained as $h(x) = \mathrm{sign}(f(x))$. The standard primal C-SVM 2-norm formulation leads to the optimization problem

$$\begin{array}{l} \min\limits_{\mathrm{w}\in\mathcal{F};b\in\mathbb{R}} \frac{1}{2}\|\mathrm{w}\|^2 + C\sum_i \xi_i \\ \text{s.t. } y_i\left(\langle \mathrm{x_i}, \mathrm{w}\rangle + b\right) + \xi_i \geq 1, \quad \xi_i \geq 0, \quad z_i \in \mathcal{Z} \end{array} \tag{2}$$

where C is a regularization term, and ξ_i are slack variables. The solution can be written as $\mathrm{w} = \sum_i \alpha_i y_i \mathrm{x_i}$, where α_i are Lagrange multipliers for the dual formulation of (2), with $\sum_i \alpha_i y_i = 0$. Vector x_i is called support vector when $\alpha_i \neq 0$. Term b is calculated a posteriori [4]. Hence, the classifier can be written as $f(x) = \sum_i \alpha_i y_i \langle \mathrm{x_i}, \mathrm{x}\rangle + b$.

For moderately unbalanced sets, empirical results show that, unlike other machine learning techniques, SVM can produce a good hypothesis, in terms of accuracy, without any modification [5, 6]. Nevertheless, performance decreases when the imbalance in the data distribution is significant. Some strategies have been proposed in the literature, which can be applied in different time during the learning process: pre-processing strategies, like over-sampling, sub-sampling, feature selection or weighted variables; training strategies, like assigning different costs through the C parameter, kernel matrix modifications; post-processing strategies, like bias tuning, and probabilistic or fuzzy output are some examples.

4 A novel post-processing strategy

Results given in the text classification domain [3] suggest a further research in post-processing strategies when learning SVM for unbalanced datasets. Furthermore, they do not directly affect the SVM training procedure.

Post-processing strategies apply on either, modifications in the weight vector of the decision function, or determining a new bias. By using this last strategy, it has been empirically shown that the learned hyperplane by a SVM in the presence of unbalanced sets has approximately the same orientation as the ideal hyperplane [7]. Hence, the critical point is the selection of the bias since standard SVM learns a boundary that is too close to the minority class: the penalty associated with errors in a small number of positive examples is exceeded by the one introduced by a large number of negative examples. Therefore, the minimization problem is more related with maximizing the margin from the examples of the majority class, resulting in a hyperplane biased towards the minority class. The so-obtained bias results in a very low generalization performance or in no generalization for patterns of this class [5].

Let $\mathcal{Z}_1 = \{z_i \in \mathcal{Z} | y_i = +1\}$, $\mathcal{Z}_2 = \{z_i \in \mathcal{Z} | y_i = -1\}$ are the sets of patterns for the positive and negative class, respectively. By defining values, $\beta = \min_{z_i \in \mathcal{Z}_1} \langle \mathrm{x_i}, \mathrm{w}\rangle$, $\alpha = \max_{z_i \in \mathcal{Z}_2} \langle \mathrm{x_i}, \mathrm{w}\rangle$ a new bias for general training datasets was defined in [4] as $b_s = -\frac{\beta+\alpha}{2}$.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Let $N_{k=1,2} = \#\mathcal{Z}_{k=1,2}$ represent the number of patterns, so an unbalanced dataset meets $N_1 \ll N_2$. Based on the proposed bias b_s, and inspired by the Fisher discriminant analysis, which takes into account the number of instances of each classes to build the decision function, a new bias $b_f = -\frac{N_1 \cdot \alpha + N_2 \cdot \beta}{N_1 + N_2}$ is defined, which reduces skew towards the minority class. Furthermore, since the SVM decision function is built based on the support vectors, the more informative instances for classification, a second bias is defined as $b_{fs} = -\frac{N_{SV_1} \cdot \alpha + N_{SV_2} \cdot \beta}{N_{SV_1} + N_{SV_2}}$, based on the same argumentation given for b_f, with $N_{SV_{k=1,2}}$ being the number of support vectors for each class.

5 Experimentation

Both new proposed post-processing strategies for a learned SVM model were experimented on twelve standard UCI datasets [8] presented in Table 1. Datasets being originally not unbalanced, they were split in the form of one class, in parenthesis in the first column, versus the rest of the classes. Performance has

Data	Total	Positive	Negative
Abalone (19)	4177	32 (0.77%)	4145 (99.23%)
Page-Blocks (5)	5473	115 (2.10%)	5358 (97.90%)
Yeast (5)	1484	51 (3.44%)	1433 (96.56%)
Car (3)	1728	69 (3.99%)	1659 (96.01%)
Ecoli (5)	336	20 (5.95%)	316 (94.05%)
Satimage (4)	6435	626 (9.73%)	5809 (90.27%)
Euthyroid	2000	238 (11.90%)	1762 (88.10%)
Glass (7)	214	29 (13.55%)	185 (86.45%)
Segmentation (1)	2310	330 (14.29%)	1980 (85.71%)
Haberman (2)	306	81 (26.47%)	225 (73.53%)
Waveform (0)	5000	1657 (33.14%)	3343 (66.86%)
Pima Diabetes (1)	768	268 (34.90%)	500 (65.10%)

Table 1: Twelve UCI datasets displayed from extreme to moderate imbalance

been evaluated on models using the Gaussian kernel, similar to [6]. The criteria used to estimate the accuracy is the 10-fold cross-validation on the whole set of training data and this procedure is repeated 3 times in order to ensure good statistical behavior. The best cross-validation g-mean rate and its standard deviation are reported in Table 2. Similarly, a sensitivity performance comparison is presented in Table 3.

Some studies can be completed according to the experimentation carried out:

- The 'general purpose' bias b_s defined in [4] generates a new decision function that ever improves the original SVM's decision function in terms of g-mean performance. It also does the same for the ten most extreme unbalanced datasets in terms of sensitivity.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Data	SVM (b)	b_s	b_f	b_{fs}
Abalone	$0.000 \pm .000$	$0.649 \pm .116$	$0.623 \pm .064$	$\mathbf{0.649 \pm .088}$
Page-Blocks	$0.530 \pm .124$	$0.718 \pm .139$	$0.828 \pm .032$	$\mathbf{0.895 \pm .028}$
Yeast	$0.000 \pm .000$	$0.646 \pm .138$	$0.641 \pm .057$	$\mathbf{0.729 \pm .125}$
Car	$0.000 \pm .000$	$0.533 \pm .169$	$\mathbf{0.936 \pm .058}$	$0.555 \pm .184$
Ecoli	$0.816 \pm .263$	$0.915 \pm .131$	$\mathbf{0.939 \pm .111}$	$0.915 \pm .131$
Satimage	$0.810 \pm .048$	$0.813 \pm .052$	$\mathbf{0.892 \pm .027}$	$0.833 \pm .047$
Euthyroid	$0.774 \pm .045$	$0.830 \pm .078$	$0.746 \pm .052$	$\mathbf{0.859 \pm .068}$
Glass	$0.000 \pm .000$	$0.858 \pm .199$	$\mathbf{0.908 \pm .068}$	$0.863 \pm .200$
Segmentation	$0.991 \pm .012$	$\mathbf{0.993 \pm .012}$	$\mathbf{0.993 \pm .012}$	$\mathbf{0.993 \pm .012}$
Haberman	$0.449 \pm .197$	$0.615 \pm .097$	$\mathbf{0.626 \pm .097}$	$\mathbf{0.626 \pm .090}$
Waveform	$0.877 \pm .013$	$0.860 \pm .019$	$\mathbf{0.883 \pm .011}$	$0.862 \pm .019$
Pima Diabetes	$0.673 \pm .065$	$0.670 \pm .065$	$\mathbf{0.735 \pm .053}$	$0.670 \pm .066$

Table 2: G-mean performance comparison for the proposed approaches

Data	SVM (b)	b_s	b_f	b_{fs}
Abalone	$0.000 \pm .000$	$0.331 \pm .149$	$\mathbf{0.906 \pm .112}$	$0.359 \pm .130$
Page-Blocks	$0.297 \pm .131$	$0.547 \pm .193$	$\mathbf{0.972 \pm .046}$	$0.958 \pm .062$
Yeast	$0.000 \pm .000$	$0.448 \pm .182$	$\mathbf{0.889 \pm .168}$	$0.571 \pm .187$
Car	$0.000 \pm .000$	$0.314 \pm .167$	$\mathbf{0.899 \pm .108}$	$0.343 \pm .190$
Ecoli	$0.733 \pm .314$	$0.867 \pm .225$	$\mathbf{0.917 \pm .190}$	$0.867 \pm .225$
Satimage	$0.677 \pm .080$	$0.678 \pm .087$	$\mathbf{0.881 \pm .061}$	$0.722 \pm .082$
Euthyroid	$0.617 \pm .075$	$0.724 \pm .138$	$\mathbf{0.957 \pm .052}$	$0.809 \pm .134$
Glass	$0.000 \pm .000$	$0.900 \pm .155$	$\mathbf{0.900 \pm .155}$	$\mathbf{0.900 \pm .155}$
Segmentation	$0.984 \pm .025$	$0.988 \pm .023$	$\mathbf{0.989 \pm .023}$	$\mathbf{0.989 \pm .023}$
Haberman	$0.269 \pm .168$	$0.545 \pm .153$	$\mathbf{0.632 \pm .160}$	$0.575 \pm .150$
Waveform	$0.824 \pm .026$	$0.776 \pm .038$	$\mathbf{0.919 \pm .018}$	$0.780 \pm .039$
Pima Diabetes	$0.510 \pm .088$	$0.498 \pm .090$	$\mathbf{0.695 \pm .088}$	$0.499 \pm .090$

Table 3: Sensitivity performance comparison for the proposed approaches

- Using the last post-processing strategy, bias b_{fs} ever improves bias b_s for both considered measures, g-mean and sensitivity.
- It is evident from Table 3 that the bias b_f moves the original bias b_s towards the majority class more than b_{fs}. Hence, the best results for the sensitivity performance are obtained. With respect to the g-mean measure, some good results are obtained. Nevertheless, several results are worse than those obtained with the bias b_s. In any case, the best mean results for both measures are provided by the bias b_f.

Therefore, it can be concluded that both Fisher-based strategies that were introduced help to improve the performance of the original SVM, as well as those using the bias b_s in its decision function, when learning from unbalanced datasets.

Finally, these results are compared with those reported in [6] where a z-SVM is presented. The z-SVM can be considered as the post-processing strategy most

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

similar to those presented in this work, because it was designed to reduce the bias of a trained SVM to the majority class for imbalanced data. It is built by reformulating the SVM's decision function as $f(x,z) = z\sum_{i:y_i=+1}\alpha_i y_i\langle \mathrm{x_i}, \mathrm{x}\rangle + \sum_{i:y_i=-1}\alpha_i y_i\langle \mathrm{x_i}, \mathrm{x}\rangle + b$. A univariate unconstrained optimization problem must be solved to determine the optimal z.

	G-mean			Sensitivity		
Data	z-SVM	b_f	b_{fs}	z-SVM	b_f	b_{fs}
Abalone	0.620	0.623	**0.649**	0.627	**0.967**	0.442
Yeast	0.728	0.641	**0.729**	0.667	**0.889**	0.571
Car	**0.936**	**0.936**	0.917	**0.917**	0.899	0.343
Euthyroid	**0.904**	0.746	0.859	0.869	**0.957**	0.809
Segmentation	0.976	**0.993**	**0.993**	0.967	**0.989**	**0.989**

Table 4: G-mean and Sensitivity performance comparison

It can be seen from Table 4 that the proposed biases obtain similar or better results than those reported using z-SVM. Moreover, the optimization problem associated with the z-SVM approach is replaced by a direct calculation of the bias.

6 Conclusions

The main contribution of this paper is that the accuracy rate of unbalanced dataset measured by geometric mean and sensitivity can be improved by using different bias. A major advantage is that the SVM optimization problem is not changed for each chosen bias and the computational cost is null. As further research, a theoretical framework to study bias movements depending on its definition is being developed.

References

[1] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, September 2009.

[2] V. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, New York, 1998.

[3] A. X. Sun, E. P. Lim, and Y. Liu. On strategies for imbalanced text classification using SVM: A comparative study. *Decision Support Systems*, 48:191–201, 2009.

[4] L. Gonzalez-Abril, C. Angulo, F. Velasco, and J. A. Ortega. A note on the bias in SVMs for multiclassification. *IEEE Transactions on Neural Networks*, 19(4):723–725, 2008.

[5] R. Akbani, S. Kwek, and N. Japkowicz. Applying support vector machines to imbalanced datasets. In *Proc. 15th European Conference on Machine Learning*, pages 39–50, 2004.

[6] T. Imam, K. Ting, and J. Kamruzzaman. z-SVM: An SVM for improved classification of imbalanced data. *AI 2006: Advances in Artificial Intelligence*, (4304):264–273, 2006.

[7] G. Wu and E. Y. Chang. KBA: Kernel Boundary Alignment Considering Imbalanced Data Distribution. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):786–795, 2005.

[8] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998.