Asymmetric Kernel Scaling for Imbalanced Data Classification

Antonio Maratea and Alfredo Petrosino

Department of Applied Science
University of Naples "Parthenope", Isola C4, Centro Direzionale, Napoli, Italy
(antonio.maratea,alfredo.petrosino)@uniparthenope.it

Abstract. Many critical application domains present issues related to imbalanced learning - classification from imbalanced data. Using conventional techniques produces biased results, as the over-represented class dominates the learning process and tend to naturally attract predictions. As a consequence, the false negative rate may result unacceptable and the chosen classifier unusable. We propose a classification procedure based on Support Vector Machine able to effectively cope with data imbalance. Using a first step approximate solution and then a suitable kernel transformation, we enlarge asymmetrically space around the class boundary, compensating data skewness. Results show that while in case of moderate imbalance the performances are comparable to standard SVM, in case of heavily skewed data the proposed approach outperforms its competitors.

1 Introduction

Data imbalance is a well known problem in the data mining community that severely biases the learning process and badly affects the performances of the mining algorithms. While both class imbalance and instance imbalance share many common aspects - and affect both the supervised than the unsupervised framework - the focus of our work is in instance imbalanced classification. Such problem is crucial in many critical application domains, for example: intrusion detection, fire alarm systems, gene selection, satellite data analysis, medical diagnosis and in all situations in which available instances of one class vastly outnumber instances of other classes. A good survey of the general problem framework may be found in [10]. The lack of data in minority class may be due to extremely expensive or impossibly uniform data gathering process, natural rarity, biased and/or partial information sources, errors, uneven sensor positioning etc. Nonetheless, rare cases are often the most interesting and valuable for the data analyst. In most real world critical applications the cost of misclassification of negative instances far exceeds the positive ones. For example in all aspects of security, the system designer surely prefers to have a reasonable amount of false alarms instead of just one undetected violation; in diagnosis of severe diseases, the doctor surely prefers to have a reasonable amount of patients wrongly alerted instead of some of them infected and undiagnosed. Leaving a classifier learn on an imbalanced dataset without corrections will produce a classification

A. Petrosino et al. (Eds.): WILF 2011, LNAI 6857, pp. 196–203, 2011.

biased towards the over-represented class and, as a side-effect, will skew the class boundary towards the under-represented class.

The aim of our paper is to propose a modification of the SVM algorithm able to balance the skewed data distribution. Nonlinear SVM are soft margin classifiers and hence our approach can be classified in the broad area of soft computing. For sake of simplicity, we consider a two class problem, but the approach is easily extensible to n classes through the classics One-Versus-One/One-Versus-All rules, or preferably through one of the more recent SVM aggregation methods proposed in literature (for example the multi class tree structured SVM proposed in [8], that seems to help by itself with unequal data distribution). Differently from many others (see related work), our approach is not based on preprocessing the data through a more or less "smart"resampling strategy: we propose an a posteriori kernel transformation that unevenly expands distances in the feature space in proximity of the boundary region. Based on real data testing with different percentages of imbalance - once parameters are well setted - the method has shown to be very effective, especially in extreme cases.

The paper is organized as follows: the most common variations of the SVM algorithm for imbalanced learning are outlined in next section, where related work is discussed; in section 3 the solution of the SVM algorithm is briefly presented and the proposed variations described in detail; in section 4 used data and experiments setup are described, results summarized; in section 5 main conclusions are drawn and future work is outlined.

2 Related Work

Many approaches have been proposed in literature to allow SVM to better cope with imbalanced data. From a general perspective, preprocessing strategies aim to mask data imbalance to the classifier and are essentially based on oversampling the minority class, undersampling the majority class or a combination of both. Strategies varies by author, some are "smarter"than others, but all share the same problem: oversampling increases artificially computational cost as adds phantom data, while undersampling may discard useful information. A popular method is Synthetic Minority Oversampling Technique [6], which rooted many variations (i.e. [1,5]).

Algorithm modifications on their side aim to bias the learning process in order to reduce majority class domination. They act on the boundary position: dynamically during learning, changing or adding some parameters or weights, or statically after learning, following proportionality criteria.

In literature the label "asymmetric Support Vector Machine"is often used to indicate Cost Sensitive Learning strategies acting on slack variables, but may broadly (and loosely) refer to any combination of the approaches previously presented. In our method, we focus on kernel modification strategies on a two class problem. We will not use data preprocessing, nor to generate, nor to discard any data. Our purpose is to modify the Kernel K in a way equivalent to asymmetrically changing the spatial resolution around the boundary.

We will not consider here the recent family of one-class classifiers, as while learning with only one target class (or only samples from one class) may be seen as an extreme case of imbalancing, data gathering and processing is actually different from learning two or more classes with heavily skewed sample distribution. In our opinion, one class classifiers should not be used when multi class labels are available.

3 Method

The classical solution of the SVM optimization problem for binary classification is the hyperplane given in following equation (see [9]):

$$f(x) = \sum_{s \in SV} \alpha_s y_s \, K(x_s, x) + b \tag{1}$$

Instance class depends on the side of the hyperplane in which each point is located, formally by the sign of $f(x)$. *Support vectors* are by definition the x_i such that $\alpha_i > 0$. If the points are linearly separable each support vector satisfies:

$$f(x_s) = y_s = \pm 1$$

When data are not linearly separable the solution may be found bounding mutipliers α_i with the condition $\alpha_i \leq C$ for - usually big - values of a positive constant C.

It has been observed that the kernel $K(x, x')$ induces a riemannian metric in the input space S through mapping ϕ [4]. The metric tensor induced by K on $x \in S$ is

$$g_{ij}(x) = \frac{\partial}{\partial x_i} \frac{\partial}{\partial x'_j} K(x, x') \mid_{x'=x} \tag{2}$$

where K is the internal product $K(x, x') = \Phi(x) \, \Phi(x')$ in some higher dimensional feature space H, where ϕ is a mapping from S to H.

The volume element with respect to this metric is:

$$dV = \sqrt{g(x)} dx_1 ... dx_n \tag{3}$$

where $g(x)$ is the determinant of the matrix whose i, j element is g_{ij}. The magnification factor $\sqrt{g(x)}$, expresses a local volume expansion under mapping ϕ.

Many kernels have been studied in literature [7] and the optimal choice of kernel is an active area of research. In the following we will consider the gaussian RBF kernel.

3.1 Conformal Kernel Transformations

A conformal transformation is a transformation that preserves local angles. To improve SVM discrimination power, authors in [2] proposed a (quasi) conformal

transformation on the kernel, with the purpose of increasing the separation between the two classes close to the boundary, so to enlarge resolution in this area. The general transformation form is:

$$K'(x, x') = D(x)K(x, x')D(x') \tag{4}$$

for a suitable definite positive function $D(x)$. If $D(x)$ and K are gaussian, then K' will also be gaussian. In this case K' satisfies Mercer condition. The specific transformation function considered in [2] follows:

$$D(x) = \sum_{x_i \in SV} e^{-k||x-x_i||^2} \tag{5}$$

where k is a positive constant. Support Vectors are by definition close to the boundary, so enlargement is applied in proximity of them, indirectly involving the boundary. The main problem is that they are not known in advance and so a two step approach is needed: first a standard SVM is needed to find an approximate solution (set of SV), then a second SVM is executed using the new transformed kernel. Such a choice can be very sensitive to SV distribution, enlarging more areas with high density of SV. Cited work does not explicitly address the data imbalance issue.

The paper [12] tries to overcome the problem of support vector distribution using a different adaptive (quasi) conformal transformation:

$$D(x) = \sum_{x_i \in SV} e^{-k_i||x-x_i||^2} \tag{6}$$

In this case each support vector has a different weight k_i that can be used to control its influence on space dilation. Considering a variable radius neighborhood for each SV and averaging the distance in feature space between the given x_i and all neighbor SV that have different labels, authors claim it is possible to calculate k_i so to compensate for irregular spatial distribution of SV. They also suggest that k_i can be used to compensate for class imbalance, assigning bigger values to the SV of the majority class.

From the same authors, an extension to non vector and non fixed dimension data is presented in [13]. The proposed transformation is:

$$D(x) = \frac{1}{|\chi_b^*|} \sum_{x_b \in \chi_b^*} e^{-k_b||\phi(x)-\phi(x_b)||^2} \tag{7}$$

where χ_b^* is the set of interpolated boundary instances and ϕ is the implicit transformation from input space S to feature space H. The first notable difference is in the way data imbalance is treated: observing that imbalanced data produce a separating hyperplane that is skewed towards the minority class, authors suggest as boundary an hyperplane found by interpolation between the center hyperplane and the majority SV hyperplane. To estimate boundary hyperplane shifting from

the center, a cost function balancing losses of false positives and false negatives is proposed. The second notable difference is that distances are computed in the feature space, instead of the input space, allowing comparison of data of different lengths and even non vectorial. The main problem of both methods [12,13] is their huge computational cost.

A similar, but more straightforward and efficient approach, called Kernel Scaling (KS) is proposed in [11]. After a preliminary standard SVM, a second one is executed with a transformed kernel based on the plain distance from the separating hyperplane, instead of support vectors. In this way points that are further from the boundary are shifted minimally, while points close to the boundary are subject to maximum shifting. The concentration of support vectors in an area is irrelevant. The proposed transformation function is:

$$D(x) = e^{-kf(x)^2} \tag{8}$$

where $f(x)$ is given by 1 and k is a positive constant independent from data. $D(x)$ reaches it maximum on the boundary surface, where $f(x) = 0$, and decays smoothly to e^{-k} on the margins, where $f(x) = \pm 1$. Data imbalance is not explicitly addressed.

3.2 Asymmetric Kernel Scaling (AKS)

The method proposed in [11] proved to be robust and efficient, but does not account for imbalanced data. We propose an extension of it that can effectively manage a markedly different number of training instances in the two classes. The basic idea is to enlarge differently areas on the two sides of the boundary surface, so to compensate for its skewness towards minority instances.

We first perform a standard SVM to compute an approximate boundary position, then we split the points in two sets, the negatives χ^- and the positives χ^+, according to first step prediction. In the second step, the applied kernel transformation function is:

$$D(x) = \begin{cases} e^{-k_1 f(x)^2}, & \text{if } x \in \chi^+ \\ e^{-k_2 f(x)^2}, & \text{if } x \in \chi^- \end{cases} \tag{9}$$

where $k_1 > k_2$ considering the positives as the majority class. In this way space enlargement is different on the two sides of the boundary surface, allowing actually to compensate the bias due to data imbalance. Classification is performed using the transformed kernel and a suitable value for k_1 and k_2. Concerning the problem of estimation of parameters, while it may seem reasonable to connect their value to the size of input data of each class, no evidence of an explicit relation emerged in the experiments. We used grid search and cross validation to find optimal values. The proposed solution is more flexible with respect to [11], includes it as a special case ($k_1 = k_2$), and proved to be effective, especially in extreme cases.

4 Data, Experiments and Results

Data used in the experiments are from a dataset downloadable from the UCI Machine Learning repository[1]. This dataset is composed of a range of biomedical voice measurements from 31 people, 23 of which are affected by Parkinson's disease (PD) [3]. Each column in the table is a voice measurement, and each row corresponds to one of 195 voice recording from these individuals. The most common learning task on these data is to discriminate healthy people from those with PD, according to "status" column, which is set to 0 for healthy and 1 for PD. To test method robustness, artificially varying percentages of skewness have been obtained, further removing minority class instances from the already imbalanced PD data. 4 block of tests have been performed: the first one considering the whole dataset; the second one removing 24 instances; the third one removing 36 instances; the fourth one removing 42 instances (extreme imbalancing), always and only from minority class. Effects of parameters is shown in figure 1.

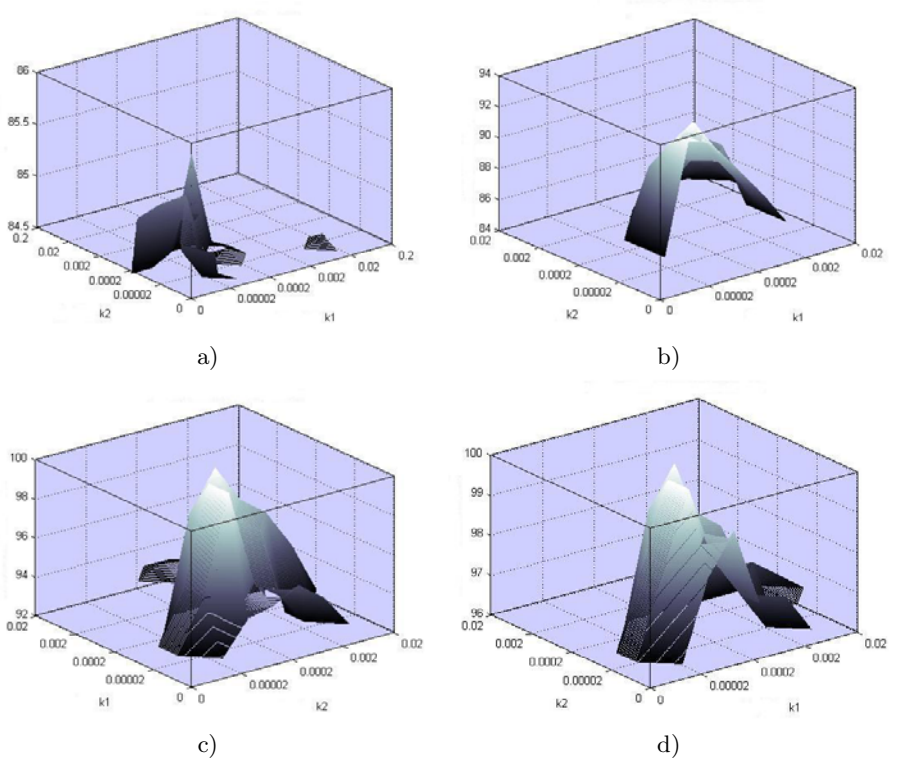

Fig. 1. Average accuracy on test sets for different sizes of the minority class a) whole dataset; b) minority class less 24 instances; c) minority class less 36 instances; d) minority class less 42 instances

[1] http://archive.ics.uci.edu/ml/datasets/Parkinsons

In the first step we perform a standard SVM classification, for which we used a gaussian kernel with base 0.5 and $C = 10$. In the second step we performed the classification with the transformed kernels. Best accuracy has been obtained with $k_1 = 2.04 \times 10^{-6}$ and $k_2 = 2 \times 10^{-5}$. These values were obtained through grid search and 5-fold cross validation.

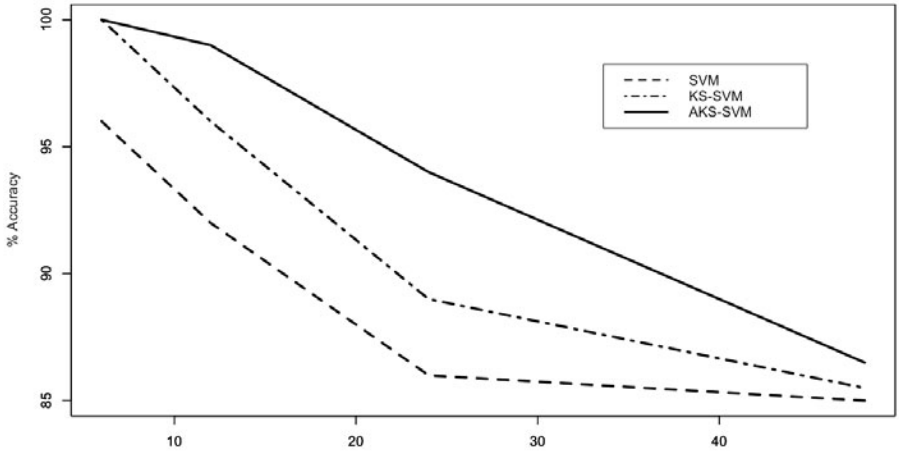

Fig. 2. X axis corresponds to the number of samples of the under-represented class, while Y axis represents the global accuracy. Classic SVM: dashed line; [11] SVM: dashed-dotted line; our algorithm: solid line.

Comparing performances with both standard SVM and the method proposed in [11], we have seen that If the data are not heavily imbalanced, there is not a remarkable advantage in using our approach, but when very few instances remain in the minority class, then our method markedly outperforms its competitors (see figure 2). We point out that the accuracy showed to be very sensitive to a good choice of parameters. Out of a narrow interval of k_1 and k_2 of effective improvement, performance tends to drop to standard SVM.

5 Conclusions

In many challenging emerging applications, like fraud detection, genetic data analysis or video classification, data are often imbalanced, and so is misclassification cost. Using conventional techniques produces biased results, as the over-represented class dominates the learning process and tend to always prevail. We presented a classification procedure based on Support Vector Machine able to effectively cope with data imbalance that is a generalization of [11]. On the basis of real medical diagnosis data, we have shown that the more the distribution is skewed, the more the proposed compensation is effective in improving the performance of the Support Vector Machine. The method has two free parameters

- whose choice critically affects performances - that have been empirically estimated. Future work is in studying a more effective estimation procedure for parameters and different kernels.

Aknowledgments. The authors thank Emanuele de Carlo for the help in code development and experiments' execution during his bachelor's degree at the University of Naples "Parthenope".

References

1. Akbani, R., Kwek, S., Japkowicz, N.: Applying support vector machines to imbalanced datasets. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) ECML 2004. LNCS (LNAI), vol. 3201, pp. 39–50. Springer, Heidelberg (2004)
2. Amari, S., Wu, S.: Improving svm classifiers by modifying kernel functions. Neural Networks 12(6), 783–789 (1999)
3. Tsanas, A., Little, M.A., McSharry, P.E., Ramig, L.O.: Accurate telemonitoring of Parkinsons disease progression by non-invasive speech tests. IEEE Transactions on Biomedical Engineering (2009) (to appear)
4. Burges, C.J.C.: Geometry and Invariance in Kernel Based Methods. In: Schlkopf, B., Burges, C.J.C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning. MIT Press, Cambridge (1999)
5. Castro, C.L., Carvalho, M.A., Braga, A.P.: An Improved Algorithm for SVMs Classification of Imbalanced Data Sets. Engineering Applications of Neural Networks, 108–118 (2009)
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Philip Kegelmeyer, W.: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
7. Cristianini, N., Shawe-Taylor, J.: An introduction to Support Vector Machines and other Kernel based Learning Methods. Cambridge University Press, Cambridge (2001)
8. Guo, J., Takahashi, N., Hu, W.: An Efficient Algorithm for Multi-class Support Vector Machines. In: International Conference on Advanced Computer Theory and Engineering (ICACTE 2008), Phuket, December 20-22, pp. 327–331 (2008)
9. Vapnik, V.: Statistical Learning Theory. John Wiley and Sons, New York (1998)
10. Weiss, G.M.: Mining with Rarity: A Unifying Framework. ACM SIGKDD Explorations 6(1), 7–19 (2004)
11. Williams, P., Li, S., Feng, J., Wu, S.: Scaling the Kernel Function to Improve Performance of the Support Vector Machine. In: Wang, J., Liao, X.-F., Yi, Z. (eds.) ISNN 2005. LNCS, vol. 3496, pp. 831–836. Springer, Heidelberg (2005)
12. Wu, G., Chang, E.Y.: Adaptive Feature-Space Conformal Transformation for Imbalanced-Data Learning. In: The Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, pp. 816–823 (2003)
13. Wu, G., Chang, E.Y.: KBA: Kernel Boundary Alignment considering imbalanced data distribution. IEEE Transaction on Knowledge and Data Engineering 17(6), 786–795 (2005)