Proceedings of International Joint Conference on Neural Networks, Montreal, Canada, July 31 - August 4, 2005

Pruning Support Vectors for Imbalanced Data Classification

Xue-wen Chen[1]*, Byron Gerlach[1], and David Casasent[2]
[1]Department of Electrical Engineering and Computer Science, The University of Kansas,
1520 West 15th Street, Lawrence, KS 66045
[2]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213

Abstract -- **In many practical applications, learning from imbalanced data poses a significant challenge that is increasingly faced by the machine learning community. The class imbalance problem raises issues that are either nonexistent or less severe compared to balanced class cases. This paper presents a new method for imbalanced data classification. The proposed method is based on support vector machine classifiers and backward pruning technique. The experimental results obtained on two data sets demonstrate the effectiveness of the new algorithm.**

I. Introduction

Imbalanced data classification refers to a two class learning problem when the number of samples in one class (typically, class of interest) is much smaller than that in the other class. Learning for imbalanced class problems is encountered in a large number of practical applications of machine learning, for example, information retrieval and filtering [1], in-flight helicopter gearbox fault monitoring [2], the detection of oil spills in satellite radar images [3], the detection of credit card fraud [4], and genomic data classification [5].

While the majority of learning methods are designed for well-balanced training data, data imbalance presents a unique challenge problem to classifier design. The class imbalance problem could hinder the performance of standard machine learning methods. For example, it is highly possible to achieve the high classification accuracy by simply classifying all samples as the class with majority samples. The problem is even severe when the misclassification costs for the two classes are different (i.e., cost-sensitive classification) and accordingly the overall classification rate is not appropriate to evaluate the performance. The practical applications of cost-sensitive classification arise frequently, for example, in medical diagnosis [6], in agricultural product inspection [7], in industrial production processes [8], and in automatic target detection [9]. Analyzing the imbalanced data thus requires new and more adaptive methods than those used in the past.

There have been a number of methods proposed to address class imbalance problems. The majority of existing methods can be grouped into two categories: sampling and weighting. In sampling-based approaches, one can either over-sample examples of the small class [10, 11] or under-sample examples of the large class [3] till the numbers of samples in both classes are approximately equal. In weighting methods, the misclassification costs of the two classes are adjusted to achieve a better performance [12, 13, 14]. Japkowicz and Stephen [15] present a thorough survey and conducted a comparative study on these methods for imbalanced data classification problems. They concluded that in most cases they studied, the weighting methods (cost-modifying) outperformed the sampling methods. The effects of data imbalance on classification systems were also evaluated in [15]. Among the three classifiers studied (C5.0 decision trees, multi-layer perceptrons (MLPs) and support vector machines (SVMs)), SVMs have been shown to be the least sensitive to the class imbalance problems [15].

In this paper, we propose a novel classification method for imbalance data problems. This method is based on support vector machines and a pruning scheme. We compare the proposed method to the SVM-based weighting method on two real world datasets. The experimental results show that the novel method yields better performance.

The paper is organized into four sections. Section II describes the SVM-based weighting method and proposed method. In Section III, we give the experimental results. Finally, conclusions are drawn in Section IV.

II. METHODS

In this session, we briefly introduce support vector machines and the cost-modifying methods for imbalanced data classification. This is followed by the proposed pruning method.

0-7803-9048-2/05/$20.00 ©2005 IEEE

A. SUPPORT VECTOR MACHINES AND COST-MODIFYING CLASSIFICATION

Consider a two-class classification problem, where the training set is described as $(y_1, \boldsymbol{x}_1), \cdots, (y_m, \boldsymbol{x}_m), \boldsymbol{x}_i \in R^n, y_i \in \{-1,+1\}$ are class labels. We define a hyperplane by the equation $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$, where $\boldsymbol{w}$ is the n-dimensional vector perpendicular to the hyperplane and b is the bias. SVMs find an optimal hyperplane that separates training samples and maximizes the margin (a margin is defined as the minimum distance between the decision surface and training samples and is shown to be $2/\|\boldsymbol{w}\|$) [16]. This is equivalent to minimizing the weight norm $\|\boldsymbol{w}\|^2$, subject to the following constraint

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) - 1 \geq 0. \qquad (1)$$

In general, a soft margin classifier is considered by introducing non-negative slack variables ξ_i to allow some training errors, as this will improve overall classification performance. The optimization problem becomes to be [17]

$$min\left(\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \xi_i\right) \qquad (2)$$

subject to

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) + \xi_i \geq 1,\ \forall i \qquad (3)$$

and

$$\xi_i \geq 0,\ \forall i \qquad (4)$$

The error tolerance for the classifier can be tuned by changing the value of C. The higher the value of C, the more errors will be allowed during training.

For the imbalanced data classification problems, a practical method is to put different penalty factors to different classes, i.e., the optimization problem in (2) becomes to be

$$min\left(\frac{1}{2}\mathbf{w}^T\mathbf{w} + C_1 \sum_{i \in class1} \xi_i + C_2 \sum_{i \in class2} \xi_i\right) \qquad (5)$$

This is referred to as cost-modifying methods. Finding best C_1 and C_2 is crucial for the performance of SVM in imbalanced data classification problems. A rule of thumb is to satisfy the ratio

$$\frac{C_1}{C_2} \propto \frac{number\ \ of\ \ samples\ \ in\ \ class2}{number\ \ of\ \ samples\ \ in\ \ class1} \qquad (6)$$

which indicates a heavier penalty on errors associated with the class of small samples. A similar formula was proposed in [18] that considers the imbalance of data and different cost of misclassification.

B. PRUNING SUPPORT VECTORS FOR IMBALANCED DATA CLASSIFICATION

We herein describe a new method for classifying imbalanced data. This method is based on training classifiers on subsets of the support vectors found for the class with majority samples. The general idea is to improve the classification rate of one class, while minimally sacrificing the rate for the other class. This is achieved by employing a pruning method to iteratively search for a subset of support vectors (from the class with larger samples) which is used to build the classification model.

For the class with larger samples, the number of support vectors is normally a small portion of the entire data. Each support vector can be considered as a representative member of a subclass, since it lies on the margin and is closest to the separating hyperplane (decision boundary) between the two classes. In fact, a number of samples may be clustered "behind" a given support vector, but it is the particular support vector which influences the decision boundary. Training a SVM classifier on the set of all support vectors only should produce the same decision boundary (hyperplane) as that yielded by using all the training samples. Removing each support vector, and then training a new classifier on the remaining support vectors, has the effect of changing the shape of the separating hyperplane, since there are less support vectors influencing that part of the decision surface. Correspondingly, removing particular support vectors of one class will cause more samples of that class to be misclassified, and in turn classify more samples as belonging to the other class.

This effect allows us to intentionally distort the decision plane by removing the influence of some groups of samples through the exclusion of their representative support vectors. Clearly, different subsets of support vectors will have different impact on the decision boundary. In conjunction with pruning methods, we can essentially find the best subset of support vectors to remove from one class (normally, the class with larger samples), which distorts the decision plane such that the classification rate for the other class improves without significantly reducing the classification rate for this class. New classifiers can be trained on a new training set, which includes all of the original class 1 samples (the class of small samples), and the selected subset of the class 2 support vectors (the class of large samples) taken from the original SVM model.

The pruning method used in this paper is backward selection. Other search strategies such as floating forward selection and branch and bound algorithms can also be used. We choose backward selection as it runs fast. Generally speaking, backward selection starts with all the support vectors (SV) and successively deletes one SV at a time. A SV is removed in an attempt to improve the classification performance for class 1 (of small samples) with minimally reduced classification rates for class 2 (of large samples). The members being removed are discarded permanently, in the order from the worst to the best, until an optimal set is found.

The proposed algorithm proceeds as follows:

```
(1) Training SVM classifies on all training samples. The support vectors of class 2 are denoted
    as $S_i = [SV_1, SV_2, \cdots, SV_m]$, where $m$ is the number of class 2 support vectors (the class with
    larger samples). Set $i = 0$.
(2) Set $n = m - i$.
For $k = 1$ to $n$
    • Exclude the $k^{th}$ support vector from $S_i$ to form a new set $S_i'$.
    • Training SVM classifiers using $S_i'$ and all the samples in class 1 as the training set.
    • Calculate the criterion function
      $J(k)$ = (class 1 accuracy)/(class 2 error).
End {for}
(3) Find g such that $J(g)$ has the smallest value among $J(k)$, $k = 1, ..., n$.
(4) Remove the $g^{th}$ SV from $S_i$.
(5) Set $i = i + 1$.
(6) Repeat steps 2 to 5 till the desired performance is achieved.
```

The process allows us to identify a subset of support vectors of class 2 which, when combined with class 1 samples, will minimize the criterion function J, i.e., by removing some support vectors of class 2, classification rate for class 1 will be improved, while the error rate of class 2 will not increase significantly.

We next apply the pruning methods for two real-world data to evaluate its performance for imbalanced data classifications.

III. EXPERIMENTAL RESULTS

The proposed method is applied to two real world datasets: Yeast dataset [19] and hyperspectral/polarimetric target detection dataset [20]. The class imbalance of Yeast dataset (about 3.5:1) is less severe than that of target detection dataset (about 25:1). Our goal is to improve the classification rate of the class with smaller samples while the error rate of the other class is not significantly increased. To evaluate the performance of the two methods (Cost-modifying and support-vector pruning), we use the confusion matrix.

A. RESULTS FOR YEAST SEQUENCE DATASET

The yeast sequence dataset consists of 1484 samples collected from SWISS-PROT using the annotations from YPD [19]. In the original dataset, proteins from yeast are classified into 10 classes, with the largest class having 463 samples and the smallest class having only 5 samples. We choose two protein classes in our experiment: membrane proteins with no N-terminal signal (ME3) and cytoskeletal (CYT). There are a total of 163 ME3 samples and 463 CYT samples, each with eight attributes. We divide the data into two sets: training and test. The training data consist of 75 ME3 and 250 CYT samples; the test data consist of 88 ME3 and 213 CYT samples.

We first train SVM classifiers using cost-modifying methods by varying weight ratios C_1/C_2 of class 1 (ME3) to class 2 (CYT) as defined in Eq. (5). Tables 2 and 3 show the confusion matrix for $C_1/C_2 = 3/1$ (as suggested in Eq. 6, close to the sample ratio) and 5/1, respectively. The results for $C_1/C_2 = 4/1$ is similar with these for $C_1/C_2 = 3/1$ and thus are not listed here. It is clear that by putting more emphasis on the errors of small class, the classification rate of ME3 is improved: when the ratio is increased from 3/1 to 4/1, seven more ME3 samples (test results) are correctly classified (this is desired) and 17 more CYT samples are misclassified.

TABLE 1: CONFUSION MATRIX FOR COST-MODIFYING METHODS WITH $C_1/C_2 = 3/1$

	Training		Test	
	ME3	CYT	ME3	CYT
ME3	68	7	**80**	8
CYT	8	242	**7**	206

TABLE 2: CONFUSION MATRIX FOR COST-MODIFYING METHODS WITH $C_1/C_2 = 5/1$

	Training		Test	
	ME3	CYT	ME3	CYT
ME3	73	2	**87**	1
CYT	33	217	**24**	189

We then run the SV-pruning method on the data. All the training samples are used to train the SVM classifier. 22 support vectors are identified for CYT class. We then iteratively remove one SV at a time. A new SVM classifier is trained based on the ME3 training samples and the remaining CYT support vectors. Tables 3 and 4 show the results for the cases when three and nine SVs are removed, respectively. As can be seen from Table 3, to correctly classify 79 ME3 samples (test), only four CYT samples are misclassified. In order to classify 87 ME3 samples, only 13 CYT samples are misclassified (see Table 4). This is better than cost-modifying methods where 24 CYT samples are misclassified (Table 2).

TABLE 3: CONFUSION MATRIX FOR SV-PRUNING METHODS WITH THREE SVS REMOVED

	Training		Test	
	ME3	CYT	ME3	CYT
ME3	69	6	**79**	9
CYT	3	247	**4**	209

TABLE 4: CONFUSION MATRIX FOR SV-PRUNING METHODS WITH NINE SVS REMOVED

	Training		Test	
	ME3	CYT	ME3	CYT
ME3	73	2	**87**	1
CYT	14	236	**13**	200

Figure 1 shows the Receiver Operating Characteristic (ROC) curve [21] for test samples, where x-axis is the false positive rate (the ratio between the number of CYT samples that are misclassified as ME3 samples and the total number of CYT samples) and y-axis is the true positive rate (the percentage of ME3 samples that are correctly classified) . As can be seen, SV-pruning methods consistently outperform cost-modifying methods: to correctly identify all the ME3 samples, the error rates of the CYT class are about 47% and 26% for cost-modifying methods and SV-pruning methods, respectively; on the other hand, to correctly identify all the CYT samples, cost-modifying methods can recognize about 45% ME3 samples only and SV-pruning methods can correctly classify 70% ME3 samples. The ROC curve allows us to determine appropriate operating point in terms of the classification performance of the two classes.

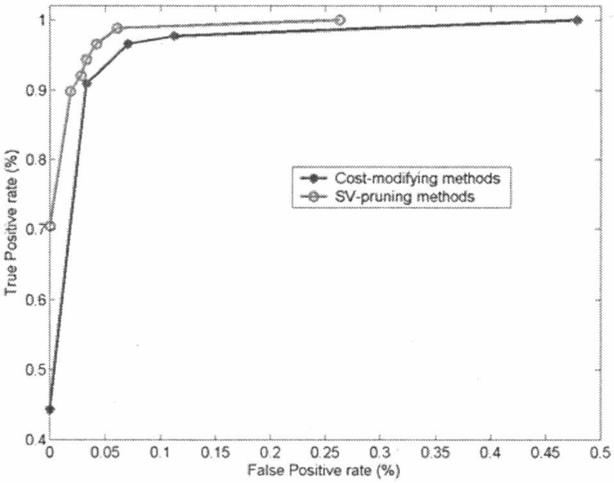

Figure 1. ROC curves for the yeast data.

B. RESULTS FOR TARGET DETECTION DATASET

The target detection dataset used here contains images with 480 by 640 pixels taken at spectral range 460 – 1000 nm with 20 nm steps at 4 different polarizations (0°, 45°, 90°, and 135°). Each scene contains one military vehicle in a vegetation background. The objective is to locate the military vehicles in the scenes. Typically, most of the pixels in the images are background pixels. We randomly extract 200 target pixels and 5000 background pixels in the hyperspectral and polarimetric set of vehicle images (HMMWV, personnel carrier, etc.). These samples are divided into two equally sized groups, one as training sample sets and the other as test set. In this paper, we report results on these training and test pixel data. The features used are the combination of HS and polarization difference image based features, as detailed in [20].

First, the SVM classifier is trained based on all the training samples. There are 83 support vectors for the background class. Table 5 shows the cost-modifying results for $C_1/C_2 = 25:1$, which is the ratio of class samples. Table 6 lists the results for SV-pruning methods by removing 48 support vectors (to achieve comparable classification rate of the target class). Using cost-modifying methods to identify 92 target samples correctly (test results), 63 background samples will be misclassified; while using SV-pruning methods, only 43 background samples are misclassified. Another interesting result is that for cost-modifying methods, if the ratio is larger than 128:1, the confusion matrix remains the same, as shown in Table 7, i.e., no further improvement can be made. Thus, the best the cost-modifying methods can do in this case is to identify 92 test samples (target) with 105 errors of background samples. For the SV-pruning methods, we can keep identifying more target samples by removing more support vectors.

TABLE 5: CONFUSION MATRIX FOR COST-MODIFYING METHODS WITH $C_1/C_2 = 25/1$

	Training		Test	
	Target	Bkg*	Target	Bkg
Target	99	1	**92**	8
Bkg	76	2424	**63**	2437

TABLE 6: CONFUSION MATRIX FOR SV-PRUNING METHODS WITH 48 SVS REMOVED

	Training		Test	
	Target	Bkg	Target	Bkg
Target	100	0	**93**	7
Bkg	33	2467	**43**	2457

TABLE 7: CONFUSION MATRIX FOR COST-MODIFYING METHODS WITH $C_1/C_2 = 128/1$

	Training		Test	
	Target	Bkg	Target	Bkg
Target	100	0	92	8
Bkg	107	2393	**107**	2393

*Bkg: Background

Figure 2 shows the Receiver Operating Characteristic curves of test samples for the HS target detection data, where x-axis is the false positive rate (the ratio between the number of background samples that are misclassified as target samples and the total number of background samples) and y-axis is the true positive rate (the percentage of target samples that are correctly classified). As can be seen, SV-pruning methods consistently outperform cost-modifying methods: to correctly identify all the target samples, about 23% background samples will be misclassified as target samples for SV-pruning methods, while cost-modifying methods can not reach 100% classification rate for the target samples (test data); on the other hand, to correctly identify all the background samples, cost-modifying methods can recognize about 20% ME3 samples only and SV-pruning methods can correctly classify 75% ME3 samples. It is worth noting that cost-modifying methods can achieve 100% classification rate of the target class for training samples, but not for test data (for test data, the best rate for the target class is 92%).

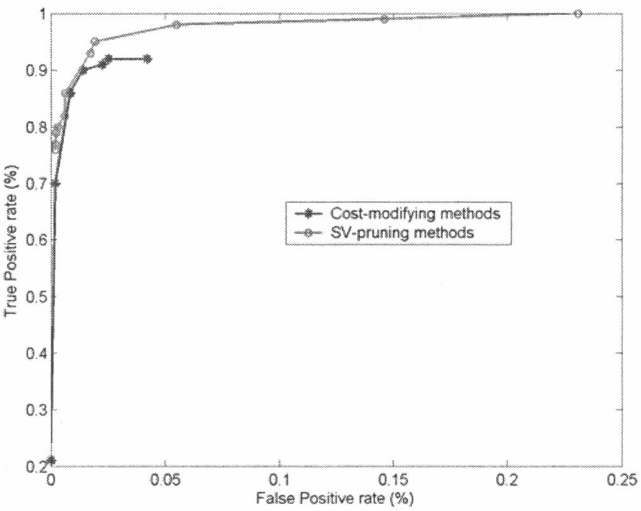

Figure 2. ROC curves for the target detection data.

IV. CONCLUSIONS

In this paper, a new method for imbalanced data classification is presented. This method is based on pruning support vectors of the class with larger samples. In support vector classifiers, support vectors are the only samples that decide the decision boundary. We first reduce the training samples of the dominant class to all the support vectors (using the support vector set and all the training samples will yield the same learning model and decision boundary). Then, by iteratively removing these support vectors that will improve the classification rate for the class with smaller samples, while minimize the effects on the class with larger samples, we can achieve a desired performance for the two classes. The method is evaluated on two real-world datasets with different levels of class imbalances and has been shown that it outperforms the cost weighting based methods. Further improvement is possible by employing other pruning approaches such as branch and bound algorithms.

ACKNOWLEDGMENT

This material is based upon work supported to Dr. Xue-wen Chen by the U. S. Army Research Laboratory and the U. S. Army Research Office under contract number DAAD19-03-1-0123. We thank Neelma Gupta of ARL for providing the HS vehicle image data used.

REFERENCES

[1]. Lewis, D. and Catlett, J., "Training text classifiers by uncertainty sampling," In *Proceedings of the Seventeenth International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 148-156, 1994.

[2]. Japkowicz, N., Myers, C., and Gluck, M., "A novelty detection approach to classification," *Proceedings of the Fourteenth Joint Conference on Artificial Intelligence*, pp. 518-523, 1995.

[3]. Kubat, M. and Matwin, S., "Learning when negative examples abound," *In Proceedings of the Ninth European Conference on Machine Learning ECML97*, pp. 179-186, 1997.

[4]. Chan, P. and Stolfo, S., "Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection," *In Knowledge Discovery and Data Mining, KDD-98*, pp. 164-168, 1998.

[5]. Craven, M., "The genomics of a signaling pathway: A KDD Cup challenge task," *SIGKDD Explorations*, 4(2), 2002.

[6]. Nunez, M., "The use of background knowledge in decision tree induction," *Machine Learning*, vol. 6, pp. 231-250, 1991.

[7]. Casasent, D. and Chen, X.-W., "New training strategies for RBF neural networks for X-ray agricultural product inspection," *Pattern Recognition*, vol. 36(2), pp. 535-547, 2003.

[8]. Verdenius, F., "A method for inductive cost optimization," *Proceedings of the Fifth European Working Session on Learning, EWSL-91*, pp. 179-191. New York: Springer-Verla, 1991.

[9]. Casasent, d. and Chen, X.-W., "Feature reduction and morphological processing for hyperspectral image data," *Applied Optics*, vol. 43 (2), pp.1-10, 2004.

[10]. Ling, C. and Li, C., "Data mining for direct marketing: Problems and solutions," *Proceeding of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)*, pp. 73-79. Menlo Park, CA: AAAI Press, 1998.

[11]. Chawla, N., Bowyer, K., Hall, L., and Kegelmeyer, P., "SMOTE: Synthetic Minority Over-sampling Technique," *International Conference on Knowledge Based Computer Systems*, 2000.

[12]. Elkan, C., "The foundations of cost-sensitive learning," *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001.

[13]. Domingos, P., "Metacost: A general method for making classifiers cost-sensitive," *Proceeding of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 155-164, 1999.

[14]. Fawcett, T. and Provost, F., "Adaptive fraud detection," *Data Mining and Knowledge Discovery*, vol. 3, pp. 291-316, 1997.

[15]. Japkowicz, N. and Stephen, S., "The Class Imbalance Problem: A Systematic Study," *Intelligent Data Analysis Journal*, vol. 6, pp. 429-449, 2002.

[16]. Cristianini, N. and Shawe-Taylor, J. *An introduction to support vector machines*, Cambridge University Press, 2000.

[17]. Vapnik, V., *Statistical learning theory*, Wiley Interscience, 1998.

[18]. Lin, Y., Lee, Y., and Wahba, G., "Support vector machines for classification in nonstandard situations," *Tech. Report, Univ. of Wisconsin*, 2000.

[19]. Horton, P. and Nakai, K., "A probabilistic classification system for predicting the cellular localization sites of proteins," *Intelligent Systems in Molecular Biology*, pp. 109-115, St. Louis, USA, 1996.

[20]. Chen, X. and Casasent, D., "Feature selection from high-dimensional hyperspectral and polarimetric data for target detection," *Proc. SPIE*, vol. 5437, pp. 171-178, 2004.

[21]. Swets, J. and Pickett, R.. *Evaluation of diagnostic systems" Methods from signal detection theory*. Academic Press, New York, NY, 1982.