International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.4 (2015), pp.105-112
http://dx.doi.org/10.14257/ijmue.2015.10.4.11

SVM Classification for High-dimensional Imbalanced Data based on SNR and Under-sampling

Li Peng[1,2], Bi Ting-ting[2] and Liu Yang[2]

[1] School of Software, Harbin University of Science and Technology, 150080 Harbin, China
[2] School of Computer Science and Technology, Harbin University of Science and Technology, 150080 Harbin, China
pli @hrbust.edu.cn.

Abstract

Support vector machine (SVM) is biased towards the majority class, in some case dataset is class-imbalanced and the bias is even larger for high-dimensional. In order to improve the classification accuracy of SVM on high-dimensional imbalanced data, we combine signal-noise ratio (SNR) and under-sampling technique based on K-means. In this article firstly we apply SNR into feature selection to reducing the feature amount then solve the problem of data imbalance using under-sampling technique based on K-means. To verify the feasibility of the proposed strategy, we utilize some metrics such as receiver operating characteristic curve (ROC curve) and area under the receiver operating characteristic curve (AUC value).As a result, the AUC value increased by 4%~16% before and after the process. The experimental results show that our strategy is feasible and effective exactly.

Keywords: *class imbalance; high-dimension; signal-noise ratio; under-sampling; SVM*

1. Introduction

At present classification technology can solve most problems and applications of some data characteristics like small amount, complete annotation and relative balance distribution. While there are still many problems limit the development of classification technology. We have to face mass data, imbalance dataset, labeling bottleneck and so on. Among them classification for imbalance dataset is one of the most challenging problems. The prediction classification performance will influenced greatly by the skewed distribution of the majority class and it proved inaccurate evaluation [1].

In fact the problem of class imbalance has drawn growing attention in machine learning [2], artificial intelligence [3] and data mining [4-5] since 2000. According to the summary of research in these areas we find it’s not satisfactory and effect of employ classifiers such as artificial neural nets (ANN) Support Vector Machine (SVM) etc directly is not ideal. Due to the special distribution of different category samples, the traditional methods could have been solved these problems appears to be inadequate, even the classification results by some methods are not be accepted. Now there are two main ways to solve class imbalance problem. One is re-sampling strategy, including over-sampling and under-sampling, to address the imbalance. The other is building a more suitable classification model, aiming at improving the classification ability of imbalance data.

Unfortunately more and more applications dataset are high-dimensional, so classification for these datasets is even larger biased in favor of the majority class [6]. In the paper we introduce a complete strategy to attenuate the deviation caused

ISSN: 1975-0080 IJMUE

by high dimension and class imbalance. Firstly, reduce dimension of dataset by feature selection method of signal-noise ratio (SNR). Secondly, cluster-based under-sampling removes samples of majority class. Besides we choose the popular SVM as our base classifiers. Promising results from experiments confirmed the effectiveness of our proposed strategy.

2. Methodology

2.1. Feature Selection

The process of feature selection is form a feature subset by selecting the representative features from the original dataset. At present, there are a lot of approaches have been put forward. Most of them are proved to be acceptable on improving the classification efficiency [7-8]. These approaches can be summarized into filter and wrapper. The former one evaluates each feature individually and assigns a score reflecting its correlation with the class label according to certain criteria [9]. While the latter one on the basis of the returned accuracy percentage of a specific classifier to search a nice solution in feature space. Generally speaking, wrapper could obtain better classification performance but expend more computational cost than filter [10]. Moreover, Golub et al. presented a much more efficient method, SNR [10]. It inherits the advantages of the above methods and is described as follow:

$$\mathrm{SNR}(i) = \left|\mu_{i1} - \mu_{i2}\right| / (s_{i1} + s_{i2}) \tag{1}$$

In the formula all samples belong to * class among them and stand for mean and standard deviation calculated. To conduct experiments, we compute SNR value for features and select top ranked features for every dataset.

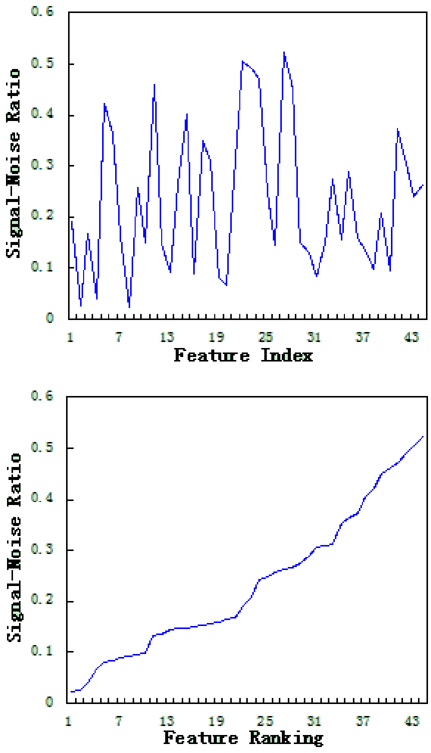


Figure 1. SNR Value Distribution for Feature Index and Feature Ranking

Take SPECTF dataset as an example, we compute SNR value for all 44 features and rank them in ascending sequence as shown in Figure 1. Many features have high SNR values and they relate to classification task closely. We select 30 features for the dataset in this study.

2.2. Under-sampling based on K-means

Clustering is one of the most common techniques for data mining. Each group formed by the clustering process is called a class. Before the clustering, it is hardly to know the number of divisional class and the data amount of each class. K-means algorithm is based on the basis principle of clustering performance index minimization. Forming a cluster according to similarity of samples and is to make samples features as similar as possible in same class moreover to ensure the relative independence from different clusters. From the above elaboration about clustering, we found much in common between clustering and hierarchy. Thus the clustering algorithm provides a good theoretical basis and feasible method for dividing layers of stratified sampling.

In the paper we apply K-means clustering in dividing layers of stratified sampling. This method is uncomplicated, effectiveness compared many typical clustering algorithms and the number of cluster K has been identified before it running. So the number of dividing layers also can be defined in advance. Therefore the sampling process is controlled effectively.

In terms of K-means algorithm, we must define the parameter K in advance, the basic idea of clustering process is described as following steps:

(1) Chose K cluster centers are randomly;

(2) To assign every sample point to the nearest cluster according to the Euclidean distance;

(3) Re-computing the center of cluster and conducting recursion based on the new one.

2.3. Support Vector Machine

Vapnik *et al.* present a classifier called support vector machine in years of research on statistics learning theory. The principles of SVM can be summarized as follow: (1) for the linear inseparable, samples are mapped into high dimensional feature space which is linear separable using nonlinear mapping algorithm; (2) It is based on the structural risk minimization theory in feature space to construct the optimal separating hyperplane for making learning machine global optimal.

Given a kernel function K and a labeled sample set, SVM finds an optimal for, in order to minimize the distance between hyperplane and the nearest sample. When inputting a test sample, the class label can be predicted by the following formula:

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{n} y_i \alpha_i K(\mathrm{x}, \mathrm{x}_i) + b \right) \tag{2}$$

$$K(x_i, x_j) = \exp\left\{ -\frac{\left| x_i - x_j \right|^2}{2\sigma^2} \right\} \tag{3}$$

Among them b is the bias of optimal classification hyperplane. In this work, the radial basis function (RBF) was selected as the kernel function.

3. The Results and Analysis of Experiment

3.1. Datasets

In this paper we select 4 datasets from UCI machine learning repository as the experimental datasets in order to verify the performance of our methods. All of them are imbalanced and have more than 30 attributes. Table 1 lists the basic information of these datasets. They are available at http://archive.ics.uci.edu/ml/datasets.html.

Table 1. The Basic Information of Four UCI Datasets

Dataset	Size	Features	Maj:Min	Imbalance ratio
SPECTF	187	44	172:15	11.47
Biodegrada	1055	41	699:356	1.96
Waveform	5000	21	3334:1666	2.00
Stu-Evaluation	5820	32	4376:1444	3.03

3.2. Evaluation Index of Experiment

The performance of the proposed methods is measured using these four datasets. All four datasets are divided into two subsets with an approximately equal number respectively and randomly. Classifier is trained and tested with them respectively.

The classification of imbalance dataset belongs to the two-classification problem. Typically majority and minority class are defined as negative and positive examples respectively in the classification task. The detail is showed as follows in Table 2. TP is the number of true negative examples; FP is the number of false positive examples; TN is the number of true negative examples; and FN is the number of false negative examples.

Table 2. Confusion Matrix

	Predicted positive class	Predicted negative class
Actual positive class	TP	FN
Actual negative class	FP	TN

In order to evaluate the effectiveness of our method we utilize ROC (receiver operating characteristic curve) and AUC (area under the receiver operating characteristic curve).

$$fpr = \frac{|FP| + 0.5}{|FP| + |TN| + 1} \tag{4}$$

$$fnr = \frac{|FN| + 0.5}{|FN| + |TP| + 1} \tag{5}$$

AUC indicates the area under receiver operating characteristic curve. It can be intuitive and clear to present classification results, the larger and the better. It is calculated as follows:

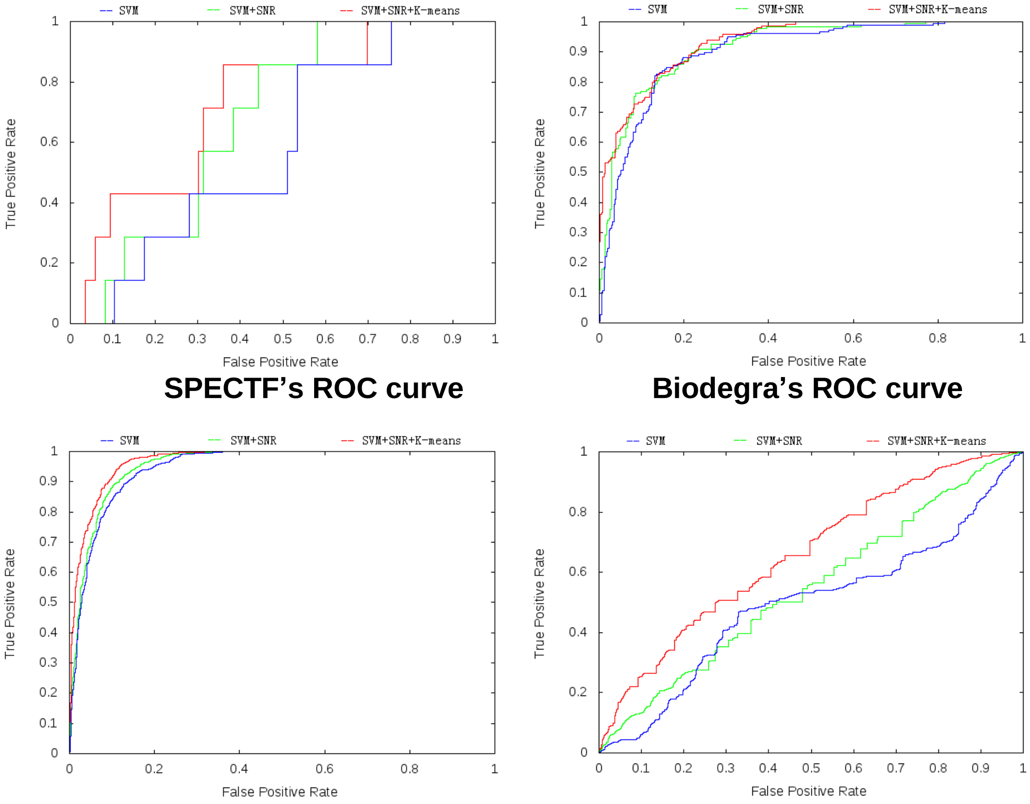
$$AUC = \int_0^1 \frac{TP}{P} d\frac{FP}{N} = \frac{1}{P \cdot N}\int_0^N TPdFP \qquad (6)$$

In the formula, $P = TP + FN$ is the number of positive and $N = TN + FP$ is the number of negative. In addition, we also introduce ROC cure that can intuitively show the classification efficiency.

3.3. Evaluation Index of Experiment

In this section we conduct experiments on four imbalanced datasets with different number of features. SVM is chose as our only classifier. To demonstrate the performance, we compare three circumstances of datasets. The first is classifying them without any pre-treatments. The second one is to reduce the dimensionality of the feature space is. At last, datasets have fewer features and are more balanced than before.

Comparative experiments based on these four datasets by our method, the ROC curve as follows:



SPECTF's ROC curve

Biodegra's ROC curve

Waveform's ROC curve

Stu-Evaluation's ROC curve

Figure 2. ROC Curve of Four UCI Data Sets

In Figure 2, the blue line represents the classification results using only SVM classifier. Classification results are respectively represented by the green line and red line after dimensionality reduction or under-sampling. On the whole, we can notice that the green lines are above the blue lines especially for the Waveform's ROC curve. In the experiment, we just keep 17 features for its samples. Actually SVM classifier behaves well on this dataset with 21 features. Then let's focus on the red lines which four datasets deal with both two proposed methods. Nearly all lines are in the red lines below. In a word, the experimental results have been reflected the advantage of our methods. As a result of ROC curve can't give quantifiable expression the level of classification performance promoting. In order to comparing in a quantitative level further, we calculated AUC values (in table 3) of these four datasets.

Table 3.The AUC Values of the Proposed Methods on Four Datasets

	SPECTF	Biodegrada	Waveform	Stu-Evaluation
SVM	0.5864	0.9010	0.9135	0.4926
SVM+SNR	0.6811	0.9164	0.9551	0.5466
SVM+SNR+K-means	0.7342	0.9574	0.9682	0.6574

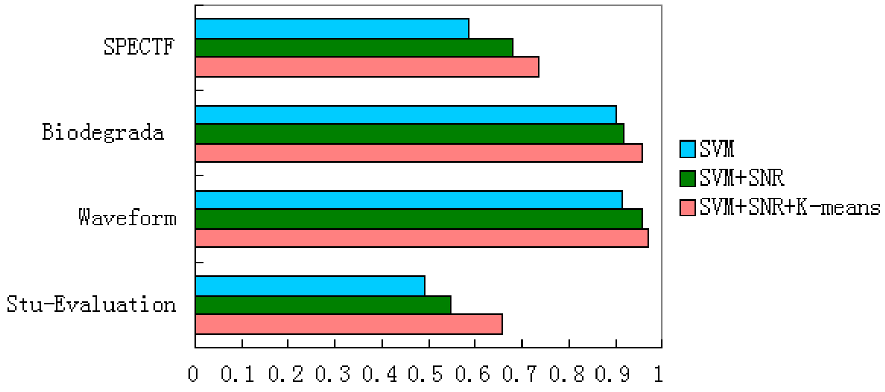


Figure 3. AUC Compare Results of Three Methods in Different Data Sets

The experiments results of three circumstances are showed in Table 3 and Figure. 3. We can see classification efficiency is improved after each treatment. To be specific, the AUC values rises by more than 5% for four datasets and by an average of 10.59%. In particular SPECTF’s and Stu-Evaluation’s rises by 14.78% and 16.48%. But for another two datasets, the AUC values have exceeded by 90% using SVM classifier only. However the values of AUC are still remain growing and classification performance is improving.

The ROC curve and AUC values explain SNR and under-sampling improve the accuracy of classification greatly in the case of high-dimensional imbalance dataset.

4. Conclusion

In this paper it’s have shown the method combine signal-noise ratio and under-sampling based on K-means we proposed has a very high efficiency to handling the high-dimensional imbalanced dataset. Experimentally verified the classification results on the four datasets of UCI based on our method is effective and it reduce the redundant features availably and properly. Moreover balance the skewed distribution simultaneously.

However there are still many big data sets in practice and some samples have many thousands of features. Worse even the large imbalance ratio exists sometimes. All of these problems will have a negative impact on the classification results. In future we intend to improve the accuracy on these problems in the light of high-dimensional imbalanced dataset. But the best thing is design a special kernel function that can solve the problem completely. So we will focus on special kernel function of imbalanced data in the future study and work.

Acknowledgements

This paper is partially supported by Foundation for University Key Teacher of Heilongjiang Province (1252G023). National Natural Science Foundation of China (61103149), China Postdoctoral Science Foundation (2011M500682, LBH-Z11106) and Province Natural Science Foundation of Heilongjiang (QC2013C060).

References

[1] L. Lusa, "Class prediction for high-dimensional class-imbalanced data [J]", BMC bioinformatics, vol. 11, no. 1, (**2010**), pp. 523-534.

[2] H. Yu, J. Ni and J. Zhao, "ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data [J]", Neurocomputing, vol. 101, no. 2, (**2013**), pp. 309-318.

[3] J. Zhang, X. Wu and V S. Sheng, "Imbalanced Multiple Noisy Labeling for Supervised Learning[C]", Proceedings of the 27th AAAI Conference on Artificial Intelligence, AAAI, (**2013**), pp. 1651-1652.

[4] S. Wang and X. Yao, "Relationships between diversity of classification ensembles and single-class performance measures [J]", IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 1, (**2013**), pp. 206-219.

[5] T R. Hoens and N V. Chawla, "Learning in non-stationary environments with class imbalance[C]", Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, (**2012**), pp. 168-176.

[6] L. Lusa, "SMOTE for high-dimensional class-imbalanced data [J]", BMC bioinformatics, vol. 14, no. 1, (**2013**), pp. 106-121.

[7] S. Sharifzadeh, L H. Clemmensen and C. Borggaard, "Supervised feature selection for linear and non-linear regression of L*a*b* color from multispectral images of meat [J]", Engineering Applications of Artificial Intelligence, vol. 27, no. 1, (**2014**), pp. 211-227.

[8] J. Rondina, T. Hahn and L. de Oliveira, "SCoRS-a method based on stability for feature selection and apping in neuroimaging [J]", IEEE Transactions on Medical Imaging, vol. 33, no. 1, (**2013**), pp. 85-98.

[9] X L. Zhang, X F Chen and Z J. He, "An ACO-based algorithm for parameter optimization of support vector machines [J]", Expert Systems with Applications, vol. 37, no. 9, (**2010**), pp. 6618-6628.

[10] T R. Golub, D K. Slonim and P. Tamayo, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring [J]", science, vol. 286, no. 11, (**1999**), pp. 531-537.