ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 132

2nd International Conference on Modelling, Simulation and Applied Mathematics (MSAM 2017)

Threshold Optimization Algorithm ε-KSVM for Unbalanced Data Classification Prediction

Guoqing Ge[1], Xin Jin[1], Xu Lu[2] and Yongbin Zhao[3]

[1]Central University of Finance and Economics, Beijing, China

[2]Inner Mongolia East Power Supply Company ICT Branch of State Grid Corporation, Inner Mongolia, China

[3]Liaoning Power Supply Company ICT Branch of State Grid Corporation, Liaoning, China

***Abstract*—In the era of Big Data, the multidimensional datasets from electric power, medical treatment and other industries are often unbalanced, and the positive data usually costs more seriously when being classified. According to different datasets, the distribution tendency of datasets may reduce the accuracy of classifiers. Based on SVM, the traditional Classifier KSVM introduces KNN algorithm effectively, that increases the effective classification information for error-prone points near the hyperplane, but at the same time it introduces more noise. Based on the defect that the KSVM algorithm with fixed threshold applied to unbalanced datasets, this paper proposes an improved ε-KSVM classifier with thresholds of dynamic adjustment for different datasets. The classifier applies Genetic Algorithm to adjust the boundary, namely the threshold dynamically so that the misclassification information is reduced. The experimental results show that the prediction accuracy is improved greatly.**

Keywords- SVM; KNN; threshold; imbalanced datasets; Genetic Algorithm

I. INTRODUCTION

With the development of network and enterprise information construction, massive data has been accumulated in power, financial, health care industry and other industries. In order to achieve core value of these massive data, data mining technology based on big data has been developed rapidly. While in practice, the unbalanced data classification problem is widespread, such as power communication fault diagnosis, medical diagnosis, network intrusion diagnosis, credit card fraud and so on. In traditional classification, maximization of classification accuracy is often based on two assumptions: 1) the number of samples in the training data set is roughly equal; 2) the cost of classification errors is roughly equal[1]. However, in the context of massive data, simplification assumption based on simplify research do no longer hold. Collecting the full amount of data become a reality. The amount of data in each category doesn't need to be artificially selected any longer. Classification problem need to be considered in a higher level overview. If using the traditional forecasting method, the data of majority class in massive data may submerge minority class, resulting that the classification of minority class is very low. And in most cases, companies are concerned with minority class data in the imbalanced data set and the losses caused by misclassification.

The misclassification cost of minority class samples is far greater than the majority's. Credit card fraud, for example, even if determining the entire samples are good credit customers in the learning process, accuracy can be as high as 90%. But such a classification is meaningless and worthless. To solve this problem, how to design the accurate classifier and improve the classification performance of a few classes is of great practical significance.

II. PREVIOUS WORK

Because of the unavoidable defects of sampling algorithm, it is a trend to solve the problem of unbalanced data from the algorithm level. The imbalance data method in algorithm level mainly includes cost-sensitive learning, single-class learning, integrated algorithm, etc[1]. Cost-sensitive learning is based on the assumption that the value of the correct classification of minority class is higher than that of the majority class. By improving the internal structure of the classifier model, the minimum error rate classifier can be transformed into a cost-sensitive classifier. In the case of severe data imbalances, single-class learning only uses the target class of interest to prevent the majority of classes from drowning by minority class. The neural network algorithm is one of the typical algorithms based on the recognition of minority class existence patterns. Integrated learning can further improve the classification performance of classifiers for minority classes by increasing the weight of the error sample classification[2]. KSVM is a relatively classic classifier, which combine SVM and KNN algorithm successfully. But the fixed threshold leading to the introduction of excessive noise is an unsolved problem. The focus of this paper lies in improving the classification prediction of unbalanced data by optimizing the KSVM classifier.

III. IMPROVEMENT OF KSVM ALGORITHM

A. KSVM Classifier

Li[3]first proposed the theorem: SVM can be seen as 1NN classifier in which each class has only one representative point. On the basis of the theorem, the KSVM algorithm is proposed. SVM takes only one representative point for each class. However, it is often not good enough to provide sufficient effective classification information. In this case, the KNN classifier is introduced, and all the support vectors are taken as representative points. Select the K (K=2n+1, n=1,2,3…) representative points of the neighbor to increase the effective classification information, and then to predict the sample class

successfully. Wang Chao, etc. pointed out the shortcomings of KSVM algorithm, and proposed an improved algorithm EDSVM [4].

The above KSVM series algorithm is based on the premise that the threshold ε is fixed, ignoring the difference of the two kinds of data information under unbalance, ignoring the number and distribution difference of the two kinds of support vectors (SVs +, SVs-) among different data sets.

B. The necessity of threshold optimization

Definition 1: The ClearData dataset is composed of the sample points to be tested, whose distance to the classifying hyperplane is greater than or equal to the threshold ε. (Figure I)

Definition 2: The BlurData dataset is composed of the sample points to be measured whose distance to the classifying hyperplane is less than the threshold ε.

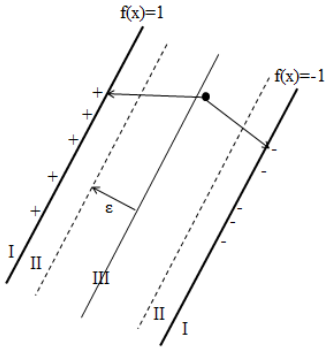

FIGURE I. HYPERPLANE AND REPRESENTATIVE POINTS

The loss function is used to estimate the degree of discrepancy between the predicted value of the model and the true value Y. It is a non-negative real-valued function. Usually use $L(Y, f(x))$ to represent it. The smaller the loss function, the better the robustness of the model. In the machine learning algorithm, the loss function hinge and SVM are inseparable.

The loss function of SVM can be expressed as

$$\text{Loss}_{\text{svm}} = \frac{1}{m}\sum_{x\in ClearData} \max\{0, 1-t\bullet y\} \tag{1}$$

The loss function of KNN can be expressed as

$$\text{Loss}_{\text{knn}} = \frac{1}{\text{n}}\sum_{x\in \text{Blur}Data} \left\|x_{\text{i}} - c_{\text{i}}\right\| \tag{2}$$

It is not difficult to obtain the loss function of KSVM classifier:

$$\text{Loss}=\text{Loss}_{\text{svm}}+\text{a}\bullet\text{Loss}_{knn}$$
$$=\frac{1}{m}\sum_{x\in ClearData} \max\{0, 1-t\bullet y\} + a\bullet\frac{1}{n}\sum_{x\in BlurData}\left\|x_i - c_i\right\| \tag{3}$$

KSVM classifier further enhances the classification accuracy by increasing the effective classification information of error-prone sample points. For the fuzzy class samples whose distance to the classification hyperplane is less than the threshold ε, the classification is carried out by KNN classifier. And the SVM classifier is adopted when the distance from the classification hyperplane is larger than the threshold ε. However, the accuracy of threshold selection has a great influence on the accuracy of the classification of samples in II and III regions, especially in minority samples.

On the one hand, when the distribution of minority class vectors is dense (Figure II), the classification results in regions II and III are inclined to majority classes. When the sample point is located near the support vector plane of minority class, the SVM computes the nearest neighbor support vector in the method of 1NN. But support vector distribution of minority class is too concentrated, which easily leading to the introduction of negative noise, so that the nearest neighbor may be the support vector of majority class (SVs-) instead. If this point is classified as a ClearData dataset, classifying it with SVM classifier will cause Losssvm increase. For this case, more efficient classification information can be obtained by increasing the K value using the algorithm KNN [5].

On the other hand, in the case that the number of minority class support vectors is significantly lower than that of majority class support vectors, the support vector distribution of minority class is too sparse (Figure II). Dense majority class of support vector provides enough information to overwhelm the classification information provided by minority class support vectors, resulting in the failure of the classifier to classify the samples in region III. At this point, if the sample points to be classified as BlurData using KNN classification, the introduction of excessive noise caused by the wrong points will increase Lossknn.

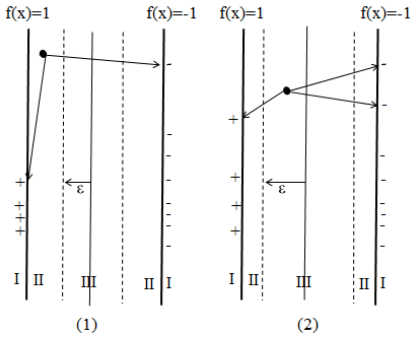

FIGURE II. EXAMPLES OF CLASSIFICATION ERROR

Therefore, increasing the effective classification of information is often accompanied by the introduction of greater noise. The boundary between ClearData and BlurData (ie, threshold ε) can be determined intelligently in this paper, which is the key to further reduce the classification "loss". The choice of threshold should be dynamically determined by the size of data sets, the number of support vectors and distribution characteristics, rather than taking a fixed threshold.

C. Propose E-KSVM Algorithm

In this paper, The main idea is that using genetic algorithm to find the optimal threshold ε* before classifying the test set with KSVM algorithm. For the clear class samples, SVM classifier is used to classify. For the BlurData class, it is necessary to increase the effective classification information. That is, using KNN classifier to classify, so as to improve the classification accuracy of the classifier samples.

ε-KSVM algorithm is as follows.

BEGIN:

Input: TrainData & TestData

Output: The classification of the samples in the test dataset

Step1: Initialize the dataset ClearData=Φ，BlurData=Φ

Step2: After learning from TrainData, get SVM classification hyperplane $g(x)=\sum_{i\in SV}\alpha_i y_i k(x_i,x)+b^*$ and support vector(SVs)

Step3: Find the optimal threshold by genetic algorithm ε*.

Step4: For the test sample, $x'\in TestData$

$$g(x')=\sum_{i\in SV}\alpha_i y_i k(x_i,x')+b^*$$

Step4.1: If $|g(x')|\geq\varepsilon^*$, ClearData=ClearData$\cup\{x'\}$;

Step4.2: If $|g(x')|<\varepsilon^*$, BlurData=BlurData$\cup\{x'\}$

Step5: TestData=TestData-x′ ;

If (TestData $\neq\Phi$), $x'\in TestData$,goto Step4.

Else go to (Step6)

Step6:

Step6.1: With regard to $x'\in ClearData$, use SVM classifier to get the classification $j^*(x')=f(x')$ of x '.

Step6.2: With regard to $x'\in$ BlurData , all SVs are used as x' neighbor samples, and KNN classifier are used to classify them.

End

The process of threshold optimization through genetic algorithm is des.

D. Threshold Optimization Process

Genetic algorithm has the implicit parallelism and strong global search ability, which can search the global optimal point in a short time. Due to the introduction of genetic operators and crossover operators, the probability of finding the global optimal solution is 1 in theory[6]. In the process of genetic algorithm threshold optimization, the fitness function guides the search direction. The purpose of this paper is to improve the classification accuracy of KSVM algorithm, so the classification accuracy (ACC) is as a sample of fitness function.

$$\text{Fitness}(\varepsilon)=\text{ACC} \tag{4}$$

Genetic algorithm is applied to KSVM threshold optimization. The basic steps of the algorithm are as follows.

BIGIN:

Step1: Initial population, and randomly select population individuals.

Step2: Each individual in the population $P(\varepsilon)$ is substituted into the fitness function $Fitness(\varepsilon)$, which is trained and tested with training data and test data[7]

Step3: According to the fitness criterion, the individual fitness function value is calculated

Step4: If the fitness function value of the optimal individual in the population is sufficiently large or the algorithm has been running for many generations, and the optimal fitness of the individual is not improved significantly, we get the optimal threshold ε *. Otherwise, continue to the next step Step5.

Step5: The selection operator is applied. According to the principle of optimal preservation and worst substitution, the next generation is selected from P (ε).

Step6: The crossover operator and the mutation operator are executed. The crossover probability is 0.7, and the mutation operator is 0.1.

End

IV. EXPERIMENTS AND DISCUSSION

TABLE I. BASIC INFORMATION OF THE DATASET

dataset	size	characteristic	Class number	distribution of treated samples	
				Train set (N+,N-)	Test set (N+,N-)
Data1	518	14	2	345 (45:300)	172 (25:147)
Data2	258	6	2	172 （32:140）	86 （19:67）
Data3	513	10	2	342 (33:309)	171 (23:148)
Data4	202	13	2	136 (18:118)	68 (22:46)
Data5	93	14	2	62 (22:40)	31 (11:20)
Data6	750	24	2	500 (50:450)	250 (25:225)
Data7	646	2	2	432 (37:395)	214 (30:184)
Data8	576	8	2	384 (34:350)	192 (19:173)

A. Data Sources

In order to verify the validity of the algorithm, eight sets of datasets with different degree of imbalance, attribute dimension and number of samples are selected from the UCI database. The data are obtained from three different industries of telecommunication, energy and medical. The basic information of the data set is shown in Table I. In this paper, F-Score and GMean are considered as classification index.

B. Experimental Results Analysis

The experiment was run on MATLAB software under Windows environment. The ε-KSVM code is written and the genetic algorithm is used to optimize the threshold. In order to avoid the influence of SVM kernel function selection, the radial basis function (RBF kernel) is chosen uniformly. σ=0.5 and the penalty factor C=2.The simulation experiment is taken by 10-fold cross validation. In this paper, we compare the advantages and disadvantages of KNN, SVM, KSVM and ε-KSVM in the evaluation criteria F-score and G-mean. KNN algorithm takes K values of 3 and 5. KSVM algorithm uses the previous experience of fixed value ε = 0.8. Then, the optimal threshold ε of eight groups of unbalanced data and the F-Score and G-mean values of the six algorithms are obtained. The optimal threshold is shown in Figure III. The results of the evaluation function are shown in Table II (Inside the brackets is the standard deviation σ of the accuracy).

From Figurea III, we can see that there is no uniform rule for the optimal thresholds of eight sets of data, and not in the vicinity of the experience threshold 0.8 taken by predecessors. But different data has different optimal threshold that needs optimization. The correctness of the ε-KSVM threshold optimization is verified by experiments. Table II shows the results of the evaluation values (in brackets is the standard deviation of multiple cross-validation). The results showed that the classification effect of KNN (K = 3) and KNN (K = 5) was similar. E-KSVM classification performance is better than KNN, KSVM, BDKSVM.

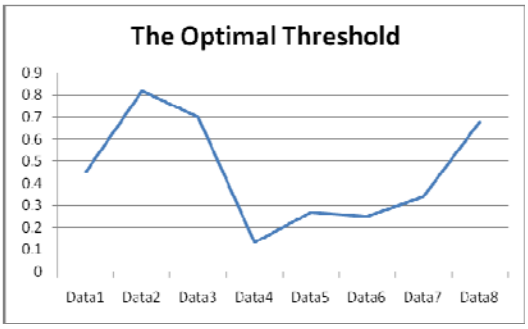

FIGURE III. THE OPTIMAL THRESHOLD

From the experimental results of G-Mean, it is not difficult to see that the improved classification accuracy of ε-KSVM is better than other algorithms. Based on the above results, it is effective to improve the classification accuracy of the unbalanced data by genetic algorithm for threshold optimization. The threshold-optimized ε-KSVM classifier proposed in this paper has a significant effect on improving the classification accuracy of unbalanced data samples, especially minority class.

TABLE II. COMPARISON OF F-SCORE AND G-MEAN

	algorithm	KNN(K=3)	KNN(K=5)	SVM	KSVM	BDKSVM	ε -KSVM
F-Score	Data1	.691(.035)	.588(.045)	.798(.054)	.766(.032)	.739(.045)	.847(.050)
	Data2	.391(.077)	.388(.064)	.439(.069)	.420(.082)	.399(.066)	.442(.059)
	Data3	.899(.063)	.901(.070)	.965(.059)	.935(.060)	.902(.080)	.972(.061)
	Data4	.690(.040)	.746(.040)	.825(.051)	.830(.045)	.792(.039)	.848(.039)
	Data5	.779(.055)	.702(.061)	.821(.056)	.871(.049)	.883(.068)	.996(.046)
	Data6	.560(.061)	.640(.071)	.920(.077)	.884(.066)	.922(.058)	.928(.051)
	Data7	.619(.071)	.677(.069)	.811(.077)	.754(.088)	.773(.081)	.812(.064)
	Data8	.721(.025)	.690(.055)	.779(.037)	.817(.054)	.822(.064)	.839(.033)
G-Mean	Data1	.527(.033)	.550(.045)	.590(.054)	.602(.036)	.594(.040)	.618(.047)
	Data2	.422(.071)	.444(.070)	.529(.080)	.532(.078)	.544(.064)	.553(.064)
	Data3	.733(.067)	.719(.072)	.774(.071)	.876(.069)	.881(.071)	.994(.065)
	Data4	.777(.044)	.739(.039)	.800(.044)	.745(.042)	.766(.043)	.824(.038)
	Data5	.781(.060)	.779(.058)	.879(.060)	.822(.059)	.866(.066)	.996(.061)
	Data6	.739(.033)	.717(.034)	.922(.042)	.899(.049)	.933(.037)	.952(.032)
	Data7	.671(.090)	.663(.088)	.772(.078)	.781(.088)	.744(.079)	.834(.078)
	Data8	.766(.033)	.733(.034)	.749(.044)	.799(.050)	.823(.035)	.827(.032)

V. CONCLUSION

In the process of data analysis of electric power, energy, medical and other industries, the problem of data classification imbalance is prevalent. In this paper, a ε-KSVM algorithm for dynamically adapting different data sets is proposed according to the distribution of unbalanced datasets. The algorithm uses genetic algorithm to find the optimal threshold of the classifier and adjusts the application range of SVM and KNN to the sample dataset, which improves the classification accuracy of

minority class, and reduces the misclassification of majority class. Compared with the previous KSVM classifiers, the ε-KSVM algorithm proposed in this paper has better performance in dealing with unbalanced classification problems. Experiments show that the algorithm is effective and scientific, and it also provides a new perspective for unbalanced binary classification. In the future, the algorithm can also be applied to power communication fault detection, disease diagnosis, credit fraud and many other practical problems.

ACKNOWLEDGEMENTS

Thanks for the following fund project's support: National Natural Science Foundation of China (U1509214) National Grid Science and Technology Project (SGTYHT/14-JS-188)

REFERENCES

[1] J. H. Wang, Intergrating KNN and Hierarchical SVM for Automatic Text Classification, Computer Applications and Software, 2016, pp. 38-41.

[2] Y. Di, Survey of Mining Imbalanced Datasets, Computer Science, 2010, pp. 27-32.

[3] R. Li, SVM-K NN Classifier-A New Method of Improving the Accuracy of SVM Classifier, Acta Electronica Sinica, 2002, pp. 745-748.

[4] C. X. Wang, Improved SVM-KNN algorithm for imbalanced datasets classification, Computer Engineering and Applications, 2016, pp. 51-55.

[5] X. Geng, Query dependent ranking using K-nearest neighbor, Annual ACM Conference on Research and Development in Information Retrieval, 2008, pp. 115-122.

[6] U. Brefeld, AUC maximizing support vector learning, Proceedings of 22nd International Conference on Machine Learning Workshop on ROC Analysis in Machine Learning, 2005.

[7] Y. Wang, Parameters Optimization of Multi-Kernel Support Vector Machine Based on Genetic Algorithm, Journal of Wuhan University, 2012, pp. 255-259.