The research of knn and svm classification performance on two kinds of unbalanced data set

DU Juan
School of Computer and Information Technology
Northeast Petroleum University
daqing, China
e-mail: dqpidj@163.com

JIANG Li-li
Qitaihe radio and television stations
Qitaihe, China
e-mail: dqpidj@163.com

***Abstract*—For Unbalanced Data Set, the KNN (K - the nearest neighbor) and SVM (support vector machine) classification algorithm's prediction result would tend to most class; the misclassification rate of the minority class was big. This paper analyzed in detail the influence of unbalanced data set to KNN and SVM in theory, and proposed a new method to solve this problem. Experiment based on UCI data set using KNN and SVM algorithm to prove the validity of the proposed method.**

Keywords- Unbalanced Data Set, classify, KNN, SVM

I. INTRODUCTION

Classification is an important research content in the field of pattern recognition and machine learning, and used in real life widely [1].Unbalanced data set classification problem is a research focus, many practical applications, such as medical diagnosis, intrusion detection, information filtering, text classification, etc. are all about unbalanced data set. The so-called unbalanced data set is that the set majority class has the big superiority than a minority class in the sample size. For the cost of unbalanced data set wrong judgment, the cost that a few class is wrong judged a majority class is more serious..

II. UNBALANCED DATA SET CLASSIFICATION PROBLEM

A. KNN Classification algorithm and unbalanced data set

The KNN classifier extends this idea by taking the k nearest points and assigning the sign of the majority [1]. KNN is a common pattern recognition algorithm, in many areas (simple and complex) shows good performance [2]. The algorithm's basic idea is, Find the K samples nearest to test sample (the most similar) in training sample set, according to the category of these samples to determine the new sample category.

Because of KNN is a typical of learning algorithm based on analogy, so each category must have a certain amount of the training samples to guarantee the accuracy of classification. In fact, such as KNN and Support Vector Machine (SVM), and other traditional classification algorithm are based on the assumptions that the training sample set is basic balance, the prediction result of classifier was biased towards the class with more samples on the data sets which contains small sample classification [3]. Figure 1 shows the influence of two kinds of sample's imbalance to the KNN classification performance. The diagram shows, the test sample belongs to the minority class, and imbalance of sample number led to the KNN classifier's discriminate error.

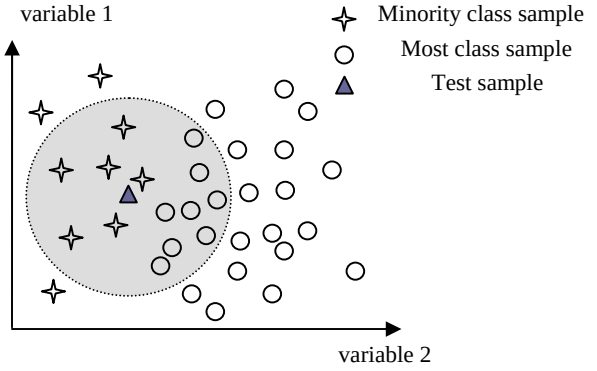

Figure 1. KNN classification on unbalanced data set

B. SVM Classification algorithm and unbalanced data set

SVM method was put forward according to the optimal classification hyperplane in the case of linear separable. It can classify all the training samples correctly, and makes the point's distance (interval) to the classification surface is the largest, that is nearest to the classification hyperplane in training sample, by interval maximization to control the complexity of the classifier, and then achieve better generalization ability. In the two kinds of pattern recognition problem, given training data, $(x_i, y_i), i = 1, \cdots, n, x \in R^d, y \in \{+1, -1\}$, SVM is refers to the classification rule (classification function) which was determined by formula(1),

$$f(x) = \operatorname{sgn}\left[\sum_{i=1}^{n} \lambda_i y_i K(x_i, x) + b\right] \tag{1}$$

$\lambda_i \geq 0, i = 1, \cdots, n$ is optimal solution of the following optimization problem,

$$\min \phi(\omega, \xi) = \frac{1}{2}\|w\|^2 + C\left[\sum_{i=1}^{l} \xi_i\right] \tag{2}$$
$$y_i(x_i \cdot w + b) \geq 1 - \xi_i, i = 1, \cdots, n \quad \xi_i \geq 0$$

Among them, ξ_i is a slack variable. C is a constant, and control the degree of punishment to wrong points sample. Using Lagrange optimization method, the optimization problem can be transformed into a dual problem, that is under the constraint conditions $\sum_{i=1}^{n}\lambda_i y_i = 0$ and $\lambda_i \geq 0$, $i = 1,\cdots,n$, to solve the maximum of the function (3) aiming at λ_i,

$$Q(\lambda) = \sum_{i=1}^{n}\lambda_i - \frac{1}{2}\sum_{i,j=1}^{n}\lambda_i\lambda_j y_i y_j \left(x_i x_j\right) \tag{3}$$

λ_i is Lagrange multiplier for each sample, this is a quadratic function optimization problem under constrained inequality, easy to prove that only a few of λ_i are not zero, these λ_i corresponding sample is support vector.

In the above analysis of the dual problem, optimal function (3) only involves inner product operation $\left(x_i \cdot x_j\right)$, this operation can be realized through the original space of the function, it is $K\left(x_i \cdot x_j\right)$ in formula (1), and now the objective function (3) can be expressed as (4).

$$Q(\lambda) = \sum_{i=1}^{n}\lambda_i - \frac{1}{2}\sum_{i,j=1}^{n}\lambda_i\lambda_j y_i y_j K\left(x_i x_j\right) \tag{4}$$

$K\left(x_i \cdot x_j\right)$ is kernel function. Commonly used kernel function are polynomial kernel function, radial basis kernel function (Gaussian kernel function) and Sigmoid kernel function, this paper used radial basis kernel function. According to the formula (3) and (4), we can get three kind of circumstance of Lagrange coefficient λ_i,

1) $\lambda_i = 0$,the sample x_i is classified correctly,

2) $0 < \lambda_i < C$, x_i is called Normal Support Vector, these Vectors are nearest to classification hyperplane in two kinds of samples, and parallel to the optimal hyperplane of training sample,

3) $\lambda_i = C$,the sample x_i is called Boundary Support Vectors, they are misclassification sample points. Obviously, boundary support vector's proportion reaction the SVM classification accuracy.

In the reference literature [2], the influence is also analyzed that the unbalanced sample data to the SVM classification accuracy, and demonstrates that when the training sample quantity tends to equilibrium, the SVM prediction tendentiousness could sharply reduce.

Below, we began to prove the influence that uneven distribution of the training samples to SVM classification accuracy.

Proof, hypothesis N_{BSV+} and N_{BSV-} is respectively the number of positive class and negative class boundary support vector, N_{SV+} and N_{SV-} is respectively the number of positive class and negative class support vector, M_+ and M_- is respectively the number of sample of positive class and negative class, according to formula (3) we can know,

$$\sum_{i=1}^{l}\lambda_i = \sum_{y_i=+1}\lambda_i + \sum_{y_i=-1}\lambda_i \tag{5}$$

$$\sum_{y_i=+1}\lambda_i = \sum_{y_i=-1}\lambda_i \tag{6}$$

Because the maximum of all λ_i is C, therefore it is,

$$N_{BSV+} \times C \leq \sum_{y_i=+1}\lambda_i \tag{7}$$

$$N_{SV+} \times C \geq \sum_{y_i=+1}\lambda_i \tag{8}$$

Synthesize (7) and (8) we can know

$$N_{BSV+} \times C \leq \sum_{y_i=+1}\lambda_i \leq N_{SV+} \times C \tag{9}$$

Similarly, we can get,

$$N_{BSV-} \times C \leq \sum_{y_i=-1}\lambda_i \leq N_{SV-} \times C \tag{10}$$

Hypothesis $\sum_{y_i=+1}\lambda_i = \sum_{y_i=-1}\lambda_i = L$, formula (9) and formula (10) were divided by and $C \times M_+$ and $C \times M_-$,we can get,

$$\frac{N_{BSV+}}{M_+} \leq \frac{L}{C \times M_+} \leq \frac{N_{SV+}}{M_+} \tag{11}$$

$$\frac{N_{BSV-}}{M_-} \leq \frac{L}{C \times M_-} \leq \frac{N_{SV-}}{M_-} \tag{12}$$

According to the formula (11) and formula (12), if $M_+ \neq M_-$ then the proportion of boundary support vector's upper bound and Lower bound in positive class and negative class, The upper bound of the minority class boundary support vector ratio is bigger than most class. This means that the misclassification proportion of minority class is bigger than most class. Such as, when the positive samples M_+ is large, $\frac{L}{C \times M_+}$ is small, $\frac{N_{BSV+}}{M_+}$ is small, namely misclassification rate is small, and vice versa [2].

C. The solving methods of unbalanced data set classification problem

The paper [4] proposed a new BPSVM algorithm; it distributes the different penalty factor for the positive and negative samples, and then obtained the better classification performance. Literature [5] proposed a intelligent methods through the sampling technology to compose minority point, and shift out redundant most point; The paper [6] proposed the over sampling algorithm based on initial classification, in literature [7], the author adopted a virtual sample space structure method based on the clustering, which can get new sample through two samples' averaging, the method is simple, but one time, each of the sample can only add a new sample, for m samples of clustering, can get $m \times (m-1)/2$ new sample at most.

This paper presented an up-sampling method based on clustering and genetic algorithm. In the optimal case three effective individuals can be generated in one time.

III. THE MINORITY CLASS SAMPLES GENERATE METHOD

Firstly, the sample space was divided into feature similarity cluster by using the K-means method, and took the samples in cluster as the parent samples, using linear crossover operator to generate new individual, then conducted mutation operation for new offspring samples with a small probability to make feature difference, avoid the offspring is near to parent excessively. Finally, verification and screening operation is done to the new individual attributes in accordance with some of the strategies, and discard invalid samples. Algorithm's specific steps are as follows,

1) Input k=5, the class C_p will be divided into 5 clusters by using K-means algorithm clustering, the i clustering is expressed as $C_i^p (1 \le i \le 5)$.

2) The individuals in the interior of C_i^p are paired at random, and then groups' number is $m_i/2$;

3) The linear crossover operator is implemented on each of the individual according to a certain probability P_{cross}, and retain all the offspring individuals and parent individuals. After a round of cross, the sample number in C_p will be $2.5m_i$, the scale of samples is 2.5 times of the original sample space.

Using the linear crossover operator, it is as shown in formula (13), a cross to produce 3 progeny individuals,

$$\begin{cases} Y_1 = 1.5X_1 - 0.5X_2 \\ Y_2 = -0.5X_1 + 1.5X_2 \\ Y_3 = (X_1 + X_2)/2 \end{cases} \tag{13}$$

4) Validation of the validity of the new individuals in C_i^p ,and discard invalid samples;

5)if the size of samples is not enough, transferring to step 2) and begin the next round of cross, until the demand quantity of samples is reached, otherwise the algorithm is end, and output C_p as the new sample space.

In this algorithm, the validation of step 4) means, controlling the quality of samples through the offspring samples' center distance.

For new samples, we expect that they have randomness and retain more information of categories, namely the samples have a good representative as far as possible, but there will still be some new samples are not suitable, and transcend the scope of clustering, we call them as invalid samples and abandoned directly. The following is the selection strategy and definition, hypothesis,

The class cluster center C_p is $C_p = \{C_1^p, C_2^p, \cdots, C_m^p\}$,

Definition 1, the radius of C_i^p is $r_i^p = \max(\left\| x_j^p - C_i^p \right\|) \quad 1 \le j \le N_p, 1 \le i \le m$,

In which N_p is the number of samples in C_i^p.

Definition 2, the distance of sample x_i to the cluster center,

$d_{ji}^p = \left\| x_j^p - C_i^p \right\| 1 \le i \le N_p, 1 \le j \le m$,

Specific validation steps are as follows,

1) Calculate the distance d_{ji}^p of new sample Y_j in C_i^p to the cluster center,

2) Take the cluster radius r_i^p as the threshold, Y_j as the effective sample which meet the conditions of $d_{ji}^p \le r_i^p$ of, put them into the training set. Otherwise they are discarded directly as invalid samples.

IV. THE EXPERIMENT AND RESULT ANALYSIS

This paper used the standard UCI data sets to test algorithm performance, and select the Breast Cancer, Vehicle, Sonar and Pima - Indian diabetes four collections of data to train and test KNN and SVM classifier. In the four data sets, only Sonar is two kinds of sample basic balance, we extract 30 samples from Sonar data set and 97 positive samples to constitute an unbalanced data set Sonar-1, So that we can get four distribution imbalanced data sets. In this data set, we selected randomly 70 percent of the sample as a training set, the rest as testing sample set for opening test.

The experiment used the original data set and the new data set which was constructed by using the method in this paper to train KNN and SVM classifier, SVM kernel was the Gaussian kernel. Experiment used the OSU - SVM MATLAB toolbox and MATLAB genetic algorithm toolbox. Final statistics 7 times the average of the experiment. Parameter selection are as follows, on DS-B, C = 2000 (penalty factor), σ = 0.01 (kernel parameter); on DS - V, C = 1200, σ = 0.06; the initial value of K is 5 of KNN classifier. When K is greater than 15, the classification accuracy is no longer occur significant change. The experiment makes $P_{cross} = 0.6$.

This paper used F - value to evaluate classification effect. The computation formula is as follows. The parameter β =2.

$$F - value = \frac{(1+\beta^2) \times \mathrm{Re}\, call \times \mathrm{Pr}\, ecision}{\beta^2 \times \mathrm{Pr}\, ecision + \mathrm{Re}\, call} \tag{14}$$

Table 3 is new sample set information which was adjusted by using the method in this paper. Figure 3 shows the SVM experimental results; Figure 4 shows the KNN experimental results.

Table 1. SVM Classifier experimental results

data set		***SVM***	
		Breast Cancer	***Vehicle***
original Data set	***minority class sample***	169	139
	majority class sample	320	453

	F-value/(%)	92.19	90.05
New Data set	***new sample***	168	293
	New minority class sample (Expansion ratio)	337(200%)	432(300%)
	majority class sample	320	453
	F-value/(%)	94.86	93.24
	F-valueImprove rate(%)	2.67	3.19

Table 2. KNN Classifier experimental results

data set		***Sonar-1***	***Pima –Indian diabetes***
originalData set	***minority class sample***	21	188
	majority class sample	78	350
	F-value/(%)	66.15	65.84
New Data set	***new sample***	55	156
	New minority class sample (Expansion ratio)	76(360%)	344(180%)
	majority class sample	78	350
	F-value/(%)	74.85	71.26
	F-value Improve rate(%)	8.70	5.42

Analyzing the experiment results we can know that the F-value is higher in the new balance data set, SVM classifier's performance is better than the KNN classifier no matter in primitive data set or the new data set , its reason is that the size of training samples is smaller, and SVM classifier can also get a better classification effect in the small sample set; KNN classifier can get better classification effect in large-scale training sample set, but this will lead to operation complexity is too high.

V. CONCLUSION

Traditional classify algorithm, the prediction result of classification was towards the majority class, when it was used to train imbalanced data sets. From the data layer sampling method angle, this paper proposed a new method to generate new samples of minority class, the clustering analysis, genetic crossover and mutation were used to structure new sample space. This paper carried out the experiment based on UCI data sets, and the results showed that in the new distribution balanced training set, KNN and SVM classifier can get better minority class classification effect. At the same time, the algorithm is completed during training stage, so it does not increase the burden of classification stage. But should notice, some parameter setting is still empirical, in addition the method only considered the sample quantity imbalance problem, in fact sample distribution can affect also classification effect, these are the problems which will be researched and solved next step. The method has important significance for some actual application that focus on minority class sample classification accuracy.

REFERENCES

[1] WU Hong-xing,PENG Yu,PENG Xi-yuan.A New Support Vector Machine Method for Unbalanced Data Treatment[J].Acta Electronica Sinica; 2006,34(12),2395-2398.

[2] Provost F. Machine leaning from imbalanced data sets[A]. Proc of 17th Nat Conf AAAI, Workshop on Imbalanced Data Sets[C].Austin, TX,2000,71-73.

[3] Xiao-yan Tao,Hong-bing Ji,Zhi-qiang Ma. the approximation SVM based on the sample distribution unbalance[J]. Computer science,2007(5) ,210-215.

[4] Chawla N, Bowyer K, Hall L, et al. Smote,synthetic minority over sampling technique[J]. Artificial Intelligence Research,2002(16),321-356.

[5] Hui Han,Lu Wang, Ming Wen,Wen-yuan Wang.over-sampling algorithm based on preliminary classification in imbalanced[J]. Computer application,2002(8) ,101-104.

[6] ZHANG Li,CHEN Gong-he.Method for Constructing Training Data Set in Intrusion Detection System[J]. Computer Engineering and Applications, 2006,42(28),145-146,180.

[7] Jun Tang. Improved K-means Clustering Algorithm Based on User Tag[J]. Journal of Convergence Information Technology. 2010,12 , 124-130.