Imbalance Data Classification Algorithm based on SVM and Clustering Function

Kai-Biao Lin
Department of Computer Science & Technology, Xiamen University of Technology, Xiamen 361024, PR China
Department of Computer Science and Engineering, Yuan Ze University Taoyuan 32026, Taiwan, ROC
kblin@xmut.edu.cn

Wei Weng
Department of Computer Science & Technology, Xiamen University of Technology, Xiamen 361024, PR China
wwweng@xmut.edu.cn

Robert K. Lai
Department of Computer Science and Engineering, Yuan Ze University Taoyuan 32026, Taiwan, ROC
krlai@cs.yzu.edu.tw

Ping Lu*
Department of Business, Xiamen University of Tec hnology, Xiamen 361024, PR China
luping@xmut.edu.cn

***Abstract*—The traditional support vector machine (SVM) was mainly used well on balanced data classification, but didn't perform well at imbalance dataset classification. In order to improve classification effects of SVM algorithm for imbalance dataset, the present paper combined the merits of FCM cluster algorithm and SVM algorithm to create a new algorithm (referred as FCM-SVM algorithm). Meanwhile, we adopted F-measure evaluation indicators, combining with predicting accuracy and recall of minority class, to evaluate algorithm classification performance. Effectiveness of FCM-SCM algorithm was verified by repeated experiences on dataset from UCI Database, the result shows that the algorithm improved the classification performance for imbalance problem compared to existing SVM algorithms.**

***Keywords*—Imbalance dataset; FCM clustering function; support vector machine**

I. INTRODUCTION

The traditional support vector machine (SVM) algorithm performed well on the classification of balance dataset. When the amount of positive class samples is almost the same as the quantity of negative class samples in a data set, it was easy to summarize significant feature of both classes. In this way, it's more accuracy and simply to defined the category of a new sample. However, many real-world datasets are imbalanced, in which most of the cases belong to a larger class and far fewer cases belong to a smaller [1-4]. If the amount of positive class samples differs greatly from the negative class in a dataset, then the feature of majority class will be much more and significant, but the feature of minority class will be very blur. Classifiers based on this kind of highly imbalance dataset will easily misclassify a new unknown minority sample to the majority class. If the imbalance dataset could be processed first, so the imbalance dataset classification problem would be transferred into balance dataset classification problem. Then SVM algorithms will be used to classifying the new transferred balance dataset. In this way, classification effects for imbalance dataset can be ensured. This paper combined the merits of FCM cluster algorithm and SVM algorithm to create a new algorithm (referred as FCM-SVM algorithm). The new algorithm improved the classification performance for imbalance dataset compared to existing SVM algorithms.

II. RELATED WORK

Approaches for addressing imbalance dataset classification problem can be divided into two main directions: sampling approaches and algorithm-based approaches [5]. Generally, sampling approaches include methods that over-sample the minority class to match the size of the majority class [6-7], and methods that under-sample the majority class to match the size of the minority class [8]. Algorithmic-based approaches are designed to improve a classifier's performance based on their inherent characteristics.

SVM algorithm was used mainly on balance dataset classification. With the extensive application of SVM, researchers developed many SVM algorithms for imbalance dataset classification. Here, we introduce some existing SVM algorithms for imbalance dataset. The first one is non-clustering normal SVM algorithm, this algorithm adopted balance dataset processing method from SVM algorithm to process imbalance dataset. It directly adopted SVM train function to establish model on training dataset of the above processed dataset. Only one classifier was developed. Then, SVM predict function and classifier were used to predict the result directly. The second one is Smote-Oversampling classification algorithm, which transferred imbalance dataset into balance dataset at first, and then used traditional SVM classification method to classify transferred dataset. When processed the training dataset, the algorithm multiplied the minority class samples until there is not much difference between the amount of minority class samples and majority class samples. Therefore the imbalance training dataset can be transferred into balance training dataset. The third one is under sampling classification algorithm, which is similar to oversampling classification algorithm. It also transferred

imbalance dataset into balance dataset firstly, then used traditional SVM algorithm to classify the transferred dataset. According to the samples amount of minority class, it randomly selected similar numbers of samples from majority class. Therefore the imbalance training dataset can be transferred into balance training dataset too. The fourth one is random classification algorithm, which used cross-validation function to obtain training dataset and testing dataset at first, and then adopt SVM train function on training dataset to build a classifier model, the classifier model and SVM predict function were used to do classification forecasting for predicting result.

III. Algorithm based on Clustering Function and SVM

In fuzzy C-means (referred as FCM) clustering function [10-11], if the sample amount radio of majority class to minority class was N, it would build N cluster centers by random. Membership matrix of each training sample and each cluster center would be automatically calculated. With Membership matrix, we can classify all the training samples to the corresponding cluster center. In this paper, we classify a training sample to the cluster center which having highest similarity in Membership matrix. The combined process of FCM function and SVM was shown in Figure 1.

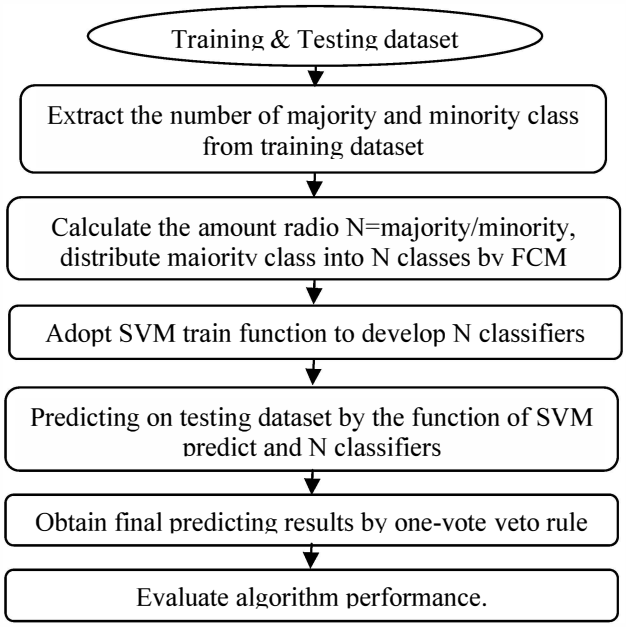

Fig. 1. Process of FCM-SVM algorithm dealing with imbalance data

The program transposed the membership matrix and extracted the corresponding cluster center index for every sample, and then recorded all the indexes in a matrix as the sample class labels; we name this matrix as idx matrix. With idx matrix, we separated majority class samples into N subclasses. After that, every subclass and the minority class comprised one new balance dataset. So, N balance datasets were formed. Then we established N classifier models by SVM train function on the new balance datasets. We used the traditional SVM classification method on every new balance dataset. One classifier model was established by SVM train function on one new balance dataset. The combinations of SVM predict function and every classifier model would be used to predict on testing dataset. The forecasting results would be well recorded at every turn.

A. Multiple Classifiers Weights Calculation

FCM clustering algorithms clustered the majority class samples into N categories. SVM train function built one model on merger dataset from each subclass and minority class. Therefore, it can get N classifiers in total. Because each sample was required to be predicted by N classifiers, so we would obtain N results for each sample. In this case, it is required to do results statistics. This paper determined final predicting result by "one vote veto" mechanism. In the classifier model, each subclass in the new balance dataset obtained from majority class only had a part of the prominent features of majority class, but minority class in the new balance dataset had all the characteristics of minority class. Therefore, if there was one classifier defined the sample belongs to majority class; the sample should belong to majority class.

The algorithm added all the predicting results of a sample together, and then divided by the amount of classifier. If the result was 1, then the sample belongs to minority class; else if the result was not 1, it means at least one predicting result was 0, so the sample belongs to majority class(i.e. the class remark of majority class is 0, and the class remark of minority is 1 for all the experiences in this paper) . The predicting was right, only when the predicting result equaled the class remark of sample in testing dataset

B. Algorithm Evaluation Measures

Despite the importance of handling imbalanced datasets, most current classification systems tend to optimize the overall accuracy without considering the relative distribution of each class. However, because predicting accuracy of the minority class affected much more than the majority class, so accuracy is no more a proper evaluation measure for imbalance problem [12-13]. Supposed that minority class comprised 1% of the total dataset, and the majority accounted for 99%. Then a simplest classifier divides all the samples into majority class, so accuracy of the classifier could achieve 99%. Obviously, this accuracy was meaningless. Because we cared more about the minority in imbalance problem, so classification accuracy was not suitable for imbalance datasets.

TABLE I. Two-class confusion matrix

	Predicted Positive	**Predicted Negative**
Actual Positive	TP(True Positive)	FN(False Negative)
Actual Negative	FP(False Positive)	TN(True Negative)

The evaluation measures used in our experiments are based on the confusion matrix. Table 1 illustrates a confusion matrix for a two class problem with positive and negative class values. With this matrix, we can derive the expression for precision and recall [14].

$$\text{Precison} = \frac{\text{TP}}{\text{TP+FP}} \qquad (1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (2)$$

The main goal for learning from imbalanced datasets is to improve the recall without hurting the precision. However, recall and precision goals can be often conflicting, since when increasing the true positive for the minority class, the number of false positives can also be increased; this will reduce the precision. F-measure combines the recall and precision on the positive class. It measures the overall performance on the minority class. The expression for the F-value is as follows:

$$\text{F} - \text{measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad (3)$$

IV. Experiment Result Analysis

A. Data Sources

All the datasets used in this paper were obtained from UCI standard database (link: http://archive.ics.uci.edu/ml/). We looked in UCI standard database and then opened proper file by dlmread or textread tools in matlab. In order to meet program requirement, we changed some structure of the dataset.

B. Experiment Process

This experiment used cross-validation function to obtain training dataset and testing dataset from pre-processed dataset at first. Then it calculated the ratio (referred as N), which approximately equaled to the amount of samples in majority class divided by the amount of samples in minority class. The majority class of training dataset remarked by class label '0' was clustered by FCM algorithm into N sub-classes. Subsequently, N sub-classes and minority class remarked by sample class label '1' were used to build N models through SVM train function. After that, predicted the class of each testing sample by every combination of classifier model and SVM predict function. Then we got the final predicting results according to 'one-vote veto' principle. The experiences also compared the class label of predicting result and the class label in testing dataset in order to get accuracy ratio, which contained majority accuracy and minority accuracy. Finally, we combined the accuracy and recall evaluating the classification performance of this algorithm, and compared the performance with other algorithms. We repeated the experiment for 10 times.

C. Experiment Result Analysis

In order to validate FCM-SVM algorithm outperformed the existing SVM algorithms, we listed the experimental results of two different types of imbalance dataset as Table 2 and Table 3, which provided an overview of the classification performance of different algorithms in diverse indexes. Each column included the different performance indexes values of one algorithm, each line showed values of a performance index for all involved algorithms. The first line showed the values of F-measure, which is the most important index in this paper; the second line was the classification accuracy values of minority class; the third line was the recall of minority class. In the Figure 2 and Figure 3, blue bar represented the most important index F-measure; red bar represented classification accuracy of minority class, and green bar represented recall of minority class.

1) Experimental Result of shuttle1 Dataset

The shuttle1 dataset was composed by two highly imbalance classes extracted from original shuttle dataset. Shuttle1 dataset was a natural imbalance dataset with two classes. We remarked the majority class, class having more samples, with class label 0, and remarked the minority class, class having fewer samples, with class label 1. The amount radio N between the majority class and minority class was 10. The experimental result was shown as Table 2 and Figure 2. The result verified FCM-SVM algorithm outperformed all the other four algorithms not only in F-measure index but also in accuracy and recall of minority.

TABLE II. Experimental Result of Shuttle1 Dataset

Evaluation Measure	F-measure	Accuracy of minority	Recall of minority
FCM-SVM	0.9232	0.8704	0.9912
Traditional SVM	0.7214	0.5891	0.8727
Smote-Oversampling	0.7951	0.9723	0.6731
Undersampling	0.7522	0.876	0.6621
Random classification	0.751	0.9823	0.6242

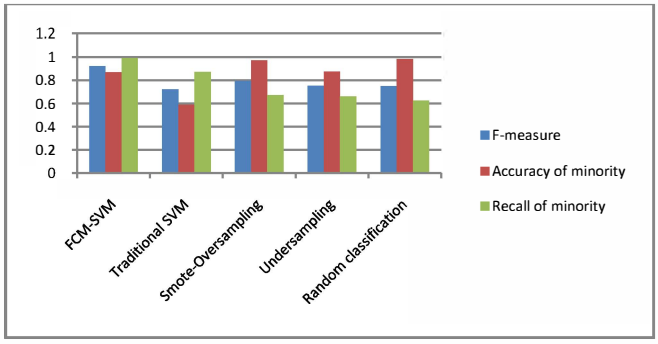

Fig. 2. Experimental result of shuttle1 dataset

2) Experimental Result of covtype1 Dataset

The dataset covtype1 was extracted from two classes of original covtype1 dataset that the original class labels were 1 and 2. It was composed as a natural imbalance dataset with two classes. We changed the majority class label's value from 1 to 0, and the minority class label's value from 2 to 1. The amount ratio of majority to minority class here was 5 to 1. The test result was shown as Table 3 and Figure 3.

TABLE III. Experimental Result of covtype1 Dataset

Evaluation Measure	F-measure	Accuracy of minority	Recall of minority
FCM-SVM	0.9521	1	0.9261
Traditional SVM	0.8847	0.9446	0.8525
Smote-Oversampling	0.8831	0.9387	0.8532

Undersampling	0.852	1	0.7752
Random classification	0.92828	1	0.87423

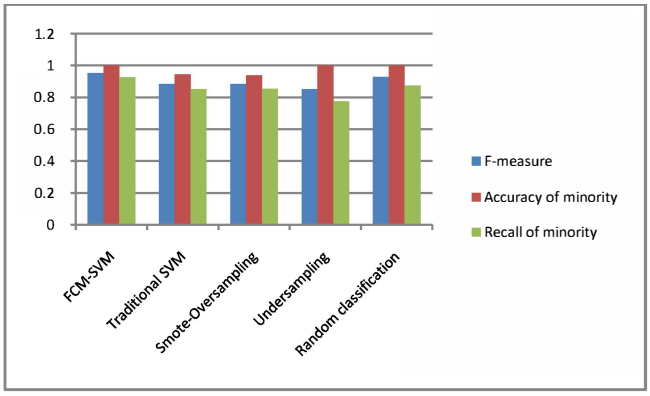

Fig. 3. Experimental result of covtype1 dataset

D. Summary of Experimental results

Experimental results showed that FCM-SVM algorithm effectively improved the classification performance for imbalance dataset compared to existing algorithms mentioned in Chapter 2. For example, in the classification prediction of shuttle1 dataset, Accuracy of minority in four SVM algorithms not combined clustering function were between 59% and 98%, but in FCM-SVM algorithm were between 86% and 92%. We cannot judge algorithms for imbalance problem only based on accuracy index. F-measure was an index combined accuracy and recall of minority and majority together. However, F-measure vales of four SVM algorithms not combined clustering function were between 70% and 79%, but the F-measure vales of FCM-SVM algorithm, proposed by this paper were between 92%-94%, as shown in Figure 4.

In the classification prediction of covtype1 dataset, classification accuracy of minority in four SVM algorithms not combined clustering function were between 94% and 100%, but in FCM-SVM algorithm was 100%. As above, we cannot judge algorithms for imbalance problem only based on accuracy index. However, F-measure vales of four SVM algorithms not combined clustering function were between 85% and 92%, but the F-measure vales of FCM-SVM algorithm, proposed by this paper was between 95%-96%, as shown in Figure 5.

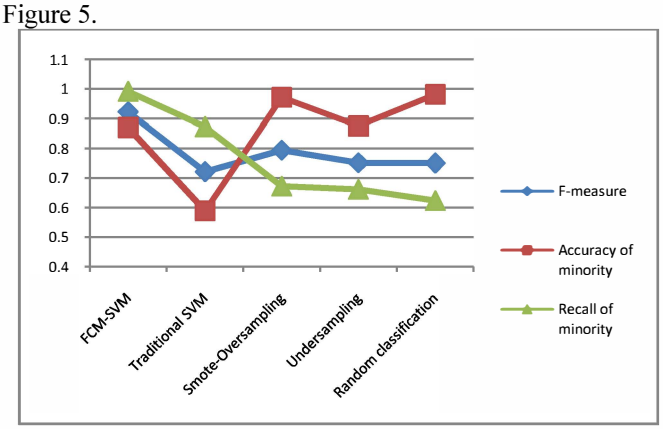

Fig. 4. Experiment result of shuttle1 dataset

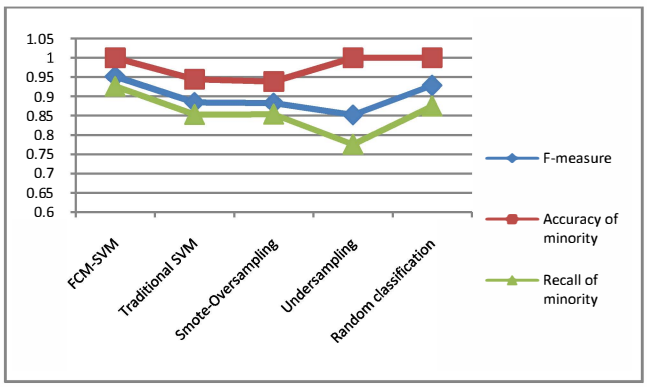

Fig. 5. Experiment result of covtype1 dataset

To sum up, the existing SVM classification algorithms kept a high classification accuracy index of minority class while the recall of minority decreased sharply. Nevertheless, FCM-SVM algorithm proposed by this paper did not only ensure the classification accuracy index of minority, but also greatly improved the recall of minority. FCM-SCM algorithm improved the classification performance for imbalance problem in general.

V. Conclusions

This paper introduced FCM clustering algorithm into the SVM classification process, in order to improve the classification performance of imbalance dataset. We can assess algorithms properly according to the algorithm evaluation measure F-measure, which combined the predicting accuracy and recall of minority class together. Experimental results illustrated FCM-SVM algorithm effectively improved classification performance for imbalance problem compared to existing SVM algorithms.

Acknowledgment

The work is partially supported by the Department of Housing and Urban-Rural Development Project under Grant No. 2013-K8-34, Xiamen University of Technology's International Cooperation and Exchange Project under Grant NO.E201300200, E201301300 and Fujian Province Department of Education Category B projects under Grant No. JB13133S.

References

[1] Jo, T. and Japkowicz, N.. Class imbalances versus small disjuncts. ACM SIGKDD Explorations Newsletter, 2004, 6(1). 40-49

[2] Dumais, S., Platt, J., Heckerman, D., and Sahami, M. Inductive Learning Algorithms and Representations for Text Categorization. In Proceedings of the Seventh International Conference on Information and Knowledge Management. 1998, pages 148–155.

[3] Ezawa, K., J., Singh, M., and Norton, S., W. Learning Goal Oriented Bayesian Networks for Telecommunications Risk Management. In Proceedings of the International Conference on Machine Learning, ICML-96, 1996. Pages 139–147, Bari, Italy. Morgan Kauffman.

[4] RADIVOJAC, P., CHAWLA, N.V., DUNKER, A.K., and OBRADOVIC, Z. Classification and knowledge discovery in protein databases. J Biomed Inform 37, 4 .2004 (Aug), 224-239. DOI= http://dx.doi.org/10.1016/j.jbi.2004.07.008.

[5] Yang Liu, Aijun An, and Xiangji Huang. Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles. W.K. Ng, M. Kitsuregawa, and J. Li (Eds.): PAKDD 2006, LNAI 3918, pp. 107–118, 2006.

[6] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[7] CHAWLA, N.V., BOWYER, K.W., HALL, L.O., KEGELMEYER, W.P. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. (JAIR), 16 2002.321–357.

[8] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou, Senior Member, IEEE.IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 39, and NO. 2, APRIL 2009. 539-550.

[9] JAMES C. BEZDEK, ROBERT EHRLICH and WILLIAM FULL.FCM: THE FUZZY c-MEANS CLUSTERING ALGORITHM. Computers & Geosciences Vol. 10, No. 2-3, pp. 1984. 191-203

[10] Xiaowei Yang, Guangquan Zhang, Jie Lu, Member, IEEE, and Jun Ma. A Kernel Fuzzy c-Means Clustering-Based Fuzzy Support Vector Machine Algorithm for Classification Problems with Outliers or Noises. IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 19, NO. 1, FEBRUARY 2011, 105-115

[11] JOSHI M V, KUMAR V, AGARWAL R C. Evaluating boosting algorithms to classify rare classes: Comparison and improvements[C] Evaluating boosting algorithms to classify rare classes: Comparison and improvements. Place Published: IEEE: 2001. 257-264.

[12] GEY, M.B.A.F., 1994. The Relationship between Recall and Precision. THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 45(1), 1994. 12-19.

[13] TANG Y, ZHANG Y Q, CHAWLA N V, et al. SVMs modeling for highly imbalanced classification [J]. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 2009, 39(1): 281-288.