A Bi-directional Sampling based on K-Means Method for Imbalance Text Classification

Jia Song, Xianglin Huang
The Faculty of Science and Technology
Communication University of China
Beijing, China
songjia@cuc.edu.cn,

Sijun Qin, Qing Song
New Media Institute
Communication University of China
Beijing, China

***Abstract*—This paper studies the imbalanced data classify-cation problem and proposes bi-directional sampling based on clustering（BDSK）for the imbalanced data classification. This algorithm combines SMOTE over-sampling algorithm and under-sampling algorithm based on K-Means to solve the within-class imbalance problem and the between-class imbalance problem. It not only avoid induce too much noise but also resolve the problem of shortage of sample. Experimental results on Tan corpus dataset show that the algorithm can effectively improve the classification performance on imbalanced data sets, especially in the cases when classification performance is heavily affected by class imbalance.**

Keywords—Data mining, Class imbalance, Text classification, K-means

I. INTRODUCTION (*HEADING 1*)

In many supervised learning applications, Imbalanced data sets are inductive learning domains in which one class is represented by a greater number of examples than the other[1].The hitch with imbalanced datasets is that standard classification learning algorithms are often biased towards the majority class (known as the "negative" class) and therefore there is a higher misclassification rate for the minority class instances (called the "positive" examples). This situation is known as the class imbalance problem[2-4]. This issue is particularly important in real- world applications where it is costly to misclassify examples from the minority class, such as diagnosis of rare diseases, spotting unreliable telecommuni-cation customers, detection of oil spills in satellite radar images, learning word pronunciations, text classification, detection of fraudulent telephone calls, information retrieval and filtering tasks, and so on[5-7]. The solutions to this problem include: (1) The data level, such as over-sampling,under-sampling and some improved methods based onthem; (2) The algorithmic level,such as cost-sensitive learning, one-sided learning and so on.

This paper combine SMOTE over-sampling algorithm and under-sampling algorithm based on K-Means to solve the within-class imbalance problem and the between-class imbalance problem.It not only avoid induce too much noise but also resolve the problem of shortage of sample.

II. THE SOLUTIONS TO IMBALANCE CLASS PROBLEM

A number of solutions to the class-imbalance problem were previously proposed to deal with this problem both at the data and algorithmic levels.

A. At the data level

Data re-sampling is used to modify the train instances in such way to product a more or less balanced class distribution that allow classifiers to preform in a similar manner to standard classification[8][9]. These solutions include many different forms of re-sampling such as random oversampling with replacement, random under sampling, directed oversampling (in which no new examples are created, but the choice of samples to replace is informed rather than random), directed under sampling (where, again, the choice of examples to eliminate is informed), oversampling with informed generation of new samples, and combinations of the above techniques[10]. A under sampling method based on K-means is proposed to keep the sample in the cluster center of majority class samples, remove redundant samples, in order to make down sampling, [11]. SMOTE[10] take the difference between the feature vector (sample) under consideration and its nearest neighbor. Multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration. This causes the selection of a random point along the line segment between two specific features. This approach effectively forces the decision region of the minority class to become more general. It can avoid classifier overfitting caused by over-sampling method based on the replication.

B. At the algorithmic level

the procedure is oriented towards the adaptation of base learning methods to be more attuned to class imbalance issues[12]. solutions include adjusting the costs of the various classes so as to counter the class imbalance, adjusting the probabilistic estimate at the tree leaf (when working with decision trees), adjusting the decision threshold, and recognition-based (i.e., learning from one class) rather than discrimination-based (two class) learning[13].

978-1-5090-0806-3/16/$31.00 copyright 2016 IEEE
ICIS 2016, June 26-29, 2016, Okayama, Japan

C. Cost-sensitive learning

This type of solutions incorporate approaches at the data level, at the algorithmic level, or at both levels combined, considering higher costs for the misclassification of examples of the positive class with respect to the negative class, and therefore, trying to minimize higher cost errors[14-15]. Mixture-of-experts approaches[12] (combining methods) have been also used to handle class-imbalance problems.

III. BI-DIRECTIONAL SAMPLING BASED ON K-MEANS (BDSK)

Class imbalance includes imbalance between classes and imbalance within a class.A between-class imbalance corresponds to the case where the number of examples representing the minority class differs from the number of examples representing the majority class; and a within-class imbalance corresponds to the case where a class is composed of a number of different clusters and the size of these clusters is also different. The within-class imbalance problem as well as the between-class imbalance problem both contribute to increasing the misclassification rate of multi-layer perceptrons [16]. In text classification, we also encountered such a problem.

In this paper, a bi-directional sampling based on K-Means (BDSK) is proposed to re-sampling on imbalance text set. Both minority and majority class also cluster use K-Means algorithm. For majority class K-Means the samples which is nearest to each cluster center are choosed to put in new dataset. The minority class is clustered into two clusters,minority cluster use SMOTE to oversample. The method can ultimately makes the balance between the various clusters, but also makes balance between classes.

A. K-Means cluster algorithm

K-Means clustering is popular for cluster analysis in data mining[17].K-Means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. The algorithm is often presented as assigning objects to the nearest cluster by distance. The main steps of K-Means are selecting the initial cluster centers, changing classification of data based on Euclidean distance and adjusting the cluster centers by classification result. The algorithm is often confused with K-Means because of the k in the name and the merits of clustering results is largely dependent on the initial cluster Select the center.

B. Bi-directional sampling based on K-Means (BDSK)

Assume that the training data set, which has multiple classes, is imbalance. The dataset contains m classes $\{C_1, C_2, ..., C_m\}$, the size of each class is $\{n_1, n_2, ..., n_m\}$.For majority class, using K-Means to partition the majority class into k clusters, where k is the arithmetic mean of classes size, choose the k nearest samples from each cluster center as a sample of new data set.For minority class, first using K-Means to partition the class into 2 clusters, choose the small cluster to do SMOTE oversampling, increase s samples to minority class. This process is repeated until the class size $ni \approx k$.

The algorithm is as follows:

- Input:Imbalance Dataset $\{C_1,C_2,...,C_m\}$, the arithmetic mean k of classes size.
- Output:New Dataset after re-sampling $\{C_1',C_2',...,C_m'\}$
- 1.for($i=0$ *to* m) do

  calculate the size of each class C_i, $n_i=sizeof(C_i)$, $i=1,2,\ldots,m$
- 2.calculate the arithmetic mean k of classes size

$$k = \sum_{i=1}^{m} n_i / m \qquad (1)$$

- 3.for(i=0 to k) do

  if (sizeof(C_i)>=k)

  C_i is majority class,randomly select k samples as the initial cluster centers, using k-Means to cluster C_i.

  Choose the k nearest samples from each cluster center as a sample of new dataset, as C_i'.

  if(sizeof(C_i)<k)

  repeat

  1. C_i is minority class,randomly select 2 samples as the initial cluster centers, using 2-Means to cluster C_i.

  2. Compare cluster.

  3.Choose the smaller cluster to SMOTE oversample, increase s samples to minority class.

  until sizeof(C_i)$\approx k$

  The new dataset has k sample,denote as C_i'

  end

This method uses K-Means to achieve majority class undersampling, while make within-class balance. Selecting the nearest sample from the cluster center, saving time from generating new samples. For minority class, this method choose smaller cluster where distribution of samples is sparse use SMOTE over-sampling, increasing the number of samples, while achieving within-class balance category. The final distribution of samples between majority class and minority class tend to balance.

The method makes training dataset balance both between-class and within-class. Over-fitting by the random over-sampling and important samples deleted by random under-sampling are avoid. The result of classifier is improved rapidly.

IV. EXPERIMENT RESULTS AND ANALYSIS

A. Performance measures

The study of this paper concentrates on imbalance problem. Since the samples between the minority classes and the majority classes are imbalance. The minority classes have much lower precision and recall than the majority classes. Accuracy places more weight on the common classes than on rare classes, which makes it difficult for a classifier to perform well on the rare classes.

By convention, the class label of the minority class is positive, and the class label of the majority class is negative. As table Ⅰ shows, *TP* and *TN* denote the number of positive and negative examples that are classified correctly, while *FN* and *FP* denote the number of misclassified positive and negative examples respectively.

TABLE I. CONFUSION TABLE

	True Positive	True Negative
Classified Positive	*TP*	*FP*
Classified Negative	*FN*	*TN*

The *Precision* for a class is the number of true positives divided by the total number of elements labeled as belonging to the positive class. The *Recall* in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class. It is defined as (2) and (3):

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The *Precision* and the *Recall* is both important and common indicator. The $F_{measure}$ considers the influence of the Precision and the Recall. It is defined as (4):

$$F_{measure} = \frac{(1+\beta^2) \times Recall \times Precision}{\beta^2 \times Recall + Precision} \tag{4}$$

In application of imbalanced dataset classification, the classification result of positives class are often more important, so $G_{measure}$ is defined to measure it. It is defined as (5):

$$G_{measure} = \sqrt{acc^+ * acc^-} \tag{5}$$

$Acc^+$and Acc^- is the accuracy of positive and negative samples respectively. They are defined as (6) and (7):

$$acc^+ = \frac{TP}{TP + FN} \tag{6}$$

$$acc^- = \frac{TN}{TN + FP} \tag{7}$$

The relationship between $G_{measure}$ and acc^+(acc^-) is nonlinear, the smaller the value of acc^+(acc^-) is, the smaller the value of $G_{measure}$ is, which means that more positives samples are classified to the wrong class. In this paper we use $F_{measure}$ (β=1,as *F1*)and $G_{measure}$ for performance measure and. We use $F_{measure}$ and $G_{measure}$ for performance measure.

B. Dataset

In order to evaluate the effectiveness of the proposed method, this paper chooses the imbalance class from Tan corpus[18]. Show as Table Ⅱ：

TABLE II. THE EXPERIMENT DATASET (PIECE)

Dataset	Sample number	Positive sample number	Negative sample number	Imbalance rate
Art	236	51	185	3.63
Hygiene	546	63	483	7.67
Finance	358	91	267	2.93
H.R.	317	28	289	10.32

C. Experiment results

We use kNN to be classification algorithm and 5-fold cross-validation and take the average of five operating results as the final result. Table Ⅲ and Ⅳ show the F_1 and G_m of using no-resampling, combine random under-sampling and SMOTE,SMOTE based on k-means, under-sampling based on k-means and BDSK in training datasets respectively.

TABLE III. *THE F1 VALUE* OF THE RE-SAMPLING

Dataset	No Re-sampling	RUS-SMOTE	SMOTE based on k-means	Under-sampling based on k-means	BDSK
Art	0.8783	0.9254	0.9476	0.9055	0.9579
Hygiene	0.8402	0.9406	0.9510	0.9311	0.9711
Finance	0.8657	0.8956	0.9217	0.8871	0.9257
H.R.	0.7905	0.9110	0.9226	0.8448	0.9675

TABLE IV. *GM* OF THE RE-SAMPLING

Dataset	No Re-sampling	RUS-SMOTE	SMOTE based on k-means	Under-sampling based on k-means	BDSK
Art	0.8165	0.9309	0.9473	0.8563	0.9581
Hygiene	0.7454	0.9387	0.9753	0.9723	0.9823
Finance	0.8283	0.8708	0.8970	0.8798	0.9250
H.R.	0.8660	0.9631	0.9817	0.8757	0.9939

Fig.1 and Fig.2 show, the *F1* value and *Gm* value comparing the kNN algorithm on no-resampling, combine random under-sampling and SMOTE,SMOTE based on k-means, under-sampling based on k-means and BDSK, the F1 value is much greater than those for common algorithm and also higher than those for random re-sampling. It is easy known that the F1 value of minority class improve significantly,the method BDSK is better than other methods.

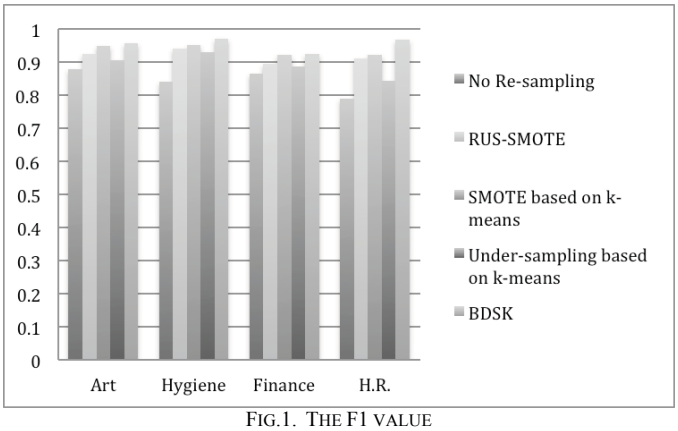

FIG.1. THE F1 VALUE

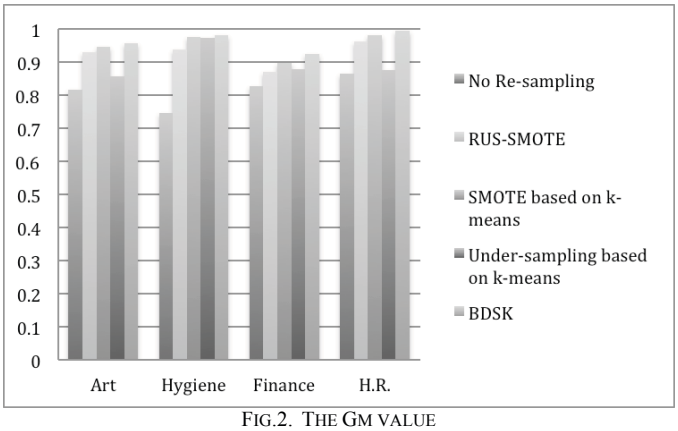

FIG.2. THE GM VALUE

V. CONCLUSION

Imbalanced data classification exists widely in real life, High imbalance data classification occurs in real-world domains; however, there is a higher misclassification rate for the minority class samples.This paper proposes the method BDSK which combine SMOTE over-sampling algorithm and under-sampling algorithm based on K-Means to make the dataset balance. Compare 5 method of re-sampling on Tan corpus data sets,the results show that the classification accuracy of majority class decline a little,this algorithm can improve the classification accuracy of minority class samples rapidly.

ACKNOWLEDGMENT

The work was supported by the project of National Key Technology R&D Program (2014BAK10B01) and project of SARFT(2014-41).

REFERENCES

[1] C.Elkan, The foundations of cost–sensitive learning, in: Proceedings of the 17th IEEE International Joint Conference on Artificial Intelligence (IJCAI'01), 2001, pp. 973-978

[2] N.V.Chawla, N.Japkowicz, A.Kotcz, Editorial: special issue on learning from imbalanced data sets, SIGKDD Explorations,2004, 6 (1) :pp.1–6.

[3] H.He, E.A.Garcia, Learning from imbalanced data, IEEE Transactions on Knowledge and Data Engineering,2009, 21 (9):pp.1263–1284.

[4] Y.Sun, A.K.C.Wong, M.S.Kamel, Classification of imbalanced data: a review, International Journal of Pattern Recognition and Artificial Intelligence, 2009, 23 (4) :pp. 687-719.

[5] J.V.Hulse, T.Khoshgoftaar.Knowledge Discovery from Imbalanced and Noisy Data[J].Knowledge and Data Engineering,2009,68(12): pp.1513-1542.

[6] Z.M.Yang, L.Y.Qiao,X.Y.Peng. Research on datamining method for imbalanced dataset based on improved SMOTE [J].ACTA Electronica Sinica, 2007, 35(12): pp.22-26.

[7] C.Drummond, R.Holte. Explicitely representing expected cost: An alternative to roc representation. In Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000, pp. 198–207.

[8] G.E.A.P.A.Batista, R.C.Prati, M.C.Monard, A study of the behavior of several methods for balancing machine learning training data, SIGKDD Explorations,2004,6 (1): pp.20–29.

[9] N.V.Chawla, K.W.Bowyer, L.O.Hall, W.P.Kegelmeyer, SMOTE: synthetic minority over-sampling technique, Journal of Artificial Intelligent Research, 2002, 16: pp. 321-357.

[10] N.V.Chawla, L.O.Hall, K.W.Bowyer, W.P.Kegelmeyer. SMOTE: synthetic minority oversampling technique. Journal of Artificial Intelligence Research, 2002, 16: pp. 321-357.

[11] S.Y.Lin, C.H.Li,Y.Jiang. Under-sampling method research in class-imbalanced data[J].Journal of Computer Research and Development,2011,48(S):pp.47-53.

[12] B.Zadrozny, C.Elkan, Learning and making decisions when costs and probabilities are both unknown, in: Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining (KDD'01), 2001, pp.204–213.

[13] F.Provost, T.Fawcett.Robust classification for imprecise environments. Machine Learning, 2001,42, pp.203-231.

[14] P.Domingos, Metacost, A general method for making classifiers cost–sensitive, in: Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining (KDD'99), 1999, pp.155–164.

[15] B.Zadrozny, J.Langford, N.Abe, Cost–sensitive learning by cost–proportionate example weighting, in: Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM'03), 2003, pp.435–442.

[16] N.Japkowicz, Concept-learning in the presence of between-class and within-class imbalances, in: Artificial Intelligence 201, LNAI 2056, pp.67–77, 2001.

[17] A.Teshansky, R.McGraw. An overview of clustering algorithms[A].Proceedings of SPIE,The International Society for Optical Engineering 1 Cj.2001(4367):41—51.

[18] S.P.Tan, Y.F.Wang. Chinese text classification corpus. [EB/OL]. TanCorp1.0. http://www.searchform.org.cn/tansongbo /corpus.htm.