SMOTE algorithm applying imbalanced data in higher education

Mengjie Zhang [1*], Jing Yang [2]

[L,2]School of computer science and technology, Guizhou University, China

[*]email:jessieyaonuli@gamil.com

Keywords: imbalanced data;smote algorithm;higher education

Abstract. In order to improve student achievement level and management ability of college , we focus on to achive the university student achievement warning system. In the practical application of College Students' performance analysis, we find that there was imbalanced data,and the proportion of the students who are dropped out of school and the students are notdropped is serious imbalanced .To solve the problem, this paper uses SMOTE algorithm to add the minority class data based the matlab platform to get balanced data, which avoids the over fitting problem that using the traditional method.

1 Introduction

Big data is maybe excessively used, but it's a real trend in today's society, International Data Corporation (IDC) defines it from the four characteristics of big data, which is the massscale data (Volume), fast data transfer and dynamic data system(Velocity), a variety of data types (Variety), a huge data value (Value) [1] .Data mining is one of important application in education of big data.Which analyses education data, mines the association between the students' performance, and enhances the students and teachers between the personalized management. The classification problem of data is the key in the process of data analysis. The traditional classification method can achieve good classification performance in solving common data set classification.However, the rapid development of computer technology has resulted in an exponential increase in the amount of data, which makes the mass data storage for data analysis, transaction management and information retrieval and so on. But what makes people interested is very limited in the amount of obtained data. It is usually only a very small part. Imabalanced data [2]is that the number of samples is far less than that of other samples.

To the classification of imbalanced data, the classifier will pay more attention to the majority of the sample, which results that in a small sample of the correct classification rate is very low. However, in practice,a small number of samples also contain more important information, and the cost of making wrong classification to samll sample is also higher. For example, To misjudge the network intrusion behavior as the normal request, it may lead to network security incidents; in some specific fields such as medical diagnosis, to mistakenly diagnose the patient suffering cancer has no cancer , it will delay the timing of treatment, which is a threat to the life of the patients .Therefore, traditional classification algorithm in the practical application is mainly to pursue high classification accuracy as the goal, which is applicable for the imbalanced data classification problem ProvostF[3] and WeissGM proved that these classification algorithm is easy to ignore minority class through experiments.

Research on data analysis in Education,the typical representative is the course of signal project in Purdue University [4], which started at the University from 2007, course signal system learners successful prediction is mainly based on the experimental data of early learning. The course signal program was started in 2007 in response to the growing decline in Rate Retention[5],which is a large proportion of the University's freshman keeping study in this school at the end of the freshman class, and the increasing graduation cycles crisis of in graduates's graduation.

The classification of multi class imbalanced data is analyzed by Wang[6] in paper. There are two main cases: a majority class and a number of minority classes, and a minority class and a number of majority classes. For a number of majoruty classes and minority classes, it can be considered as the first two cases occuring simultaneously.

2 SMOTE algorithm

SMOTE algorithm is proposed in the Journal of artificial intelligence in the first time by Chawla[7] in 2002. It is a over-sampling method. The main idea of this method is to insert a new sample between two minority classes of samples which are close in distance with k-nearest neighbor and linear interpolation accordance with a certain rules ,which can achieve the purpose of increasing the minority class of samples and make the data balanced.The SMOTE over-sampling algorithm is a new type over-sampling technique which is different from the traditional sampling algorithm. The traditional sampling method is add into the original data set by copying some sample simply, however the SMOTE technology uses the way producing a new minority class samples to change the distribution of the data set, to avoid a large number of repeated samples in data sets, and to reduce the degree of imbalance data. We can see from the new sample characteristics of the SMOTE technology , it can solve the over-fitting problem which appears esaily to a certain degree.

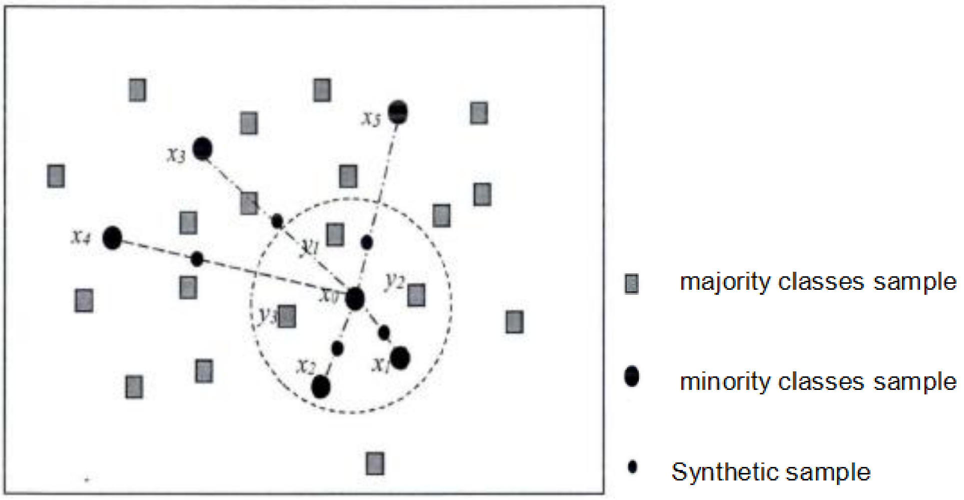

Fig.1 Basic principal of SMOTE method

The process of synthesizing a new sample by SMOTE[8]algorithm is shown in Figure 1. We regard the minority class sample X_O as an example, firstly calculate the similar K- nearest neighbor (k=5) samples which is made into{X_1~X_5}. We can see from the figure, the minority class X_0 near the sample is surrounded by heterogeneous samples excessively, which also is to use the traditional classification algorithm to solve the imbalanced data classification problem can not have a good effect.

And then randomly select a sample from the 5- adjacent to the same sample, assuming X_i ,then calculate the difference between the sample X_0 and the X_i attribute value, and the corresponding D-value from X_0 and i,diffi= X_{1i}-X_{0i} ,in which, X1i is the i attribute value of the sample X_1. We can assume that the data set has n-dimensional attribute, and then according to the calculation of the 1.1 type,using D-value diffi multiply by a random number from [0,1], then plus the corresponding attribute value Xlifrom the X_1 sample ,and you can generate a new attribute value f_{0i} .For the minority class sample X_0, each dimension can obtain by a new attribute value, which can be combined into a new minority class sample f_0 according with the sequence of corresponding dimension.

$$f_{oi} = X_{oi} + \text{diffi } X*\text{rand}[0,1] \qquad (1)$$

Then, according to sampling rate which is set in advance, the process is repeated over and over again, and a new sample is synthesized,adding it to minority class sample sets,which can get a new sample as a new training set.It can be seen from the graph ,the essence of this technique is that a new sample is randomly inserted into the line formed the current sample and its random sample of a k-adjacent sample. The new sample using this method can expand the distribution space of minority classes, making that the classifier trained on the new training set has better generalization ability and classification ability.

3 Methods and Analysis

As we know ,it’s important to complete their studies in universtity for sudents, to study influential factors causing students to be dropped out of school is key parts in college education . when analysing the data, it is found that the number of people who are dropped out of school and who are not differs extremely ,For a example, a academic department in our school, there are 1930 data records in the two semester, inlcuding 113 records of who are dropped out and 1817 records of who are not. Obviously, the proportion of data is seriously imbalanced.

If we classify this imbalanced data , the classifier will pay more attention to the majority sample that students who are not dropped out, which will lead to the correct rate of classification that minority samples is very low. So We must solve the imbalanced data firstly, then we analyze fully data and mining beteen the elements,to achieve a students score warning system , so as to better manage student scores, mobilize the enthusiasm of students learning, to avoid being dropped out of school.

In this case, we use the software matlab2015b to achieve the SMOTE algorithm. Here are the main steps of the program.

(1)To wite SMOTEZ program in the editor area (P.S: Remember to set the path,saving program in the bin folder below the matlab file ,),part of the code is as follows,

```
if(nargin < 2) ; help smote;
   if(k>NT-1)
      k=NT-1;
     warning('not so many instances in T.k is set to %d',k);
```

(2) import data. Import student information ziguan2013.xls by the format matrix numeric.Select import selection,Then you can see the workspace has emerged to the range of data you just choosed.

(3)SMOTE algorithm. Input code as follow in window command

```
   DemoOut=SMOTEZ (ziguan2013', 1800);
```

And then execute this code, selection evaluate.

(4)Click demoOut in workspace .

We’ll see the data has been processed by SMOTE algorithm. After the final data is balanced, the following figure:

Table 1 Description of original data and balanced data

Sample	Dropped out of school or not		Total
	YES(1)	NO(0)	
Original training sample	113	1817	1930
Balanced traning sample	1913	1917	3830

(P.S:YES(1) reprensents students who are dropped out of school,NO(0) represents students who are not dropped out))

And then we use the F-value criteria that is one of evaluation principle about imbalanced data set classification ,it is defined as:

$$F\text{-value}=2*\frac{recall*precision}{recall+precison} \tag{2}$$

$$recall=\frac{\mathrm{TP}}{(TP+FN)} \qquad precision=\frac{\mathrm{TP}}{(TP+FP)} \tag{3}$$

Fianlly,We use SPSS software to analyze and predict the data that has been balanced. Then get:
Precision=1819/ (1819+95) =95% ; Recall=1819/ (1819+273) =87%;

Then F-value=2RP/ (R+P) =91%

We can see from the results, after the SMOTE algorithm,the data is balanced, and the F-value value was very good.

4 Summary

Imblanced data is common seen, but it has not been paid attention to. Before the SMOTE algorithm, the random sampling method is used to deal with the imbalanced data,which randomly copys minority classes samples, so it has certain blindness and limitations. However the SMOTE algorithm is based on the theory of linear interpolation, according to certain rules to the sample for the purpose of artificial synthesis.Therefore, the SMOTE algorithm is very good to avoid the information redundancy of minority classes samples, and improve the classification performance of the data set.

Based on the SMOTE algorithm, the proposed SMOTEBoost algorithm, in the literature[9], is a combination of SMOTE and Boosting algorithm. In the literature[10],propose to the combination of SMOTE and Biased-SVM algorithm. These algorithms based on the SMOTE algorithm can solve the problem of different kinds of imbalanced data.

Acknowledgment

This paper is sponsored by National Natural Science Foundation of China (61262029, 61462013，NSFC61562011).

Finally,sincerely thanks for my advisor, Chih-Yueh Chou ,Shu-Fen Tseng and my classmate Kung-Chen Peng from Yuan Ze university in TaiWan ,they save time in the busy teaching work to modify and check my paper. Thank you all very much.

References

[1]PENG D,DABEK F.Large-scale incremental processing using distributed transaction and notifications[C].Berkeley,CA:USENIX Association,2010:1-15.

[2]Weiss GM. Mining with rarity: A unifying framework[J]. SIGKDD Explorations,2004, 6(1): 7-19.

[3] Weiss G M,Provost F J, Learning when training data are costly: The effect of class distribution on tree induction[J]. J. Artif. Intell. Res.(JAIR), 2009,19: 315-354.

[4]ARNOLD, K.E, PISTILLI, M.D.Course signals at purdue: u-sing learning analytics toIncrease student succese[C].New York:ACM Press,2012.

[5]Wang S,Yao X. Multiclass imbalance problems: Analysis and potential solutions[J], Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 2012, 42(4): 1119-1130.

[6]PISTILLI,M.D, ARNOLEK.E.Purdue signals: mining real-time academic data to enhance student success[J].About Campus:Ecriching the student learning experience,2010(3).

[7]Chawla NV, Bowyer KW , Hall LO,et al. Smote: synthetic minority over-sampling technique[J], Journal of Artificial Intelligence Research,2002,16:321-357

[8]Bo Zhou,Research and Application of Imblanced Data Classification Algorithms Based on Ensemble Learning.DALIAN UNIVERSITY OF TECHNOLOGY.2014

[9]Chawla N.V., Lazarevic A., Hall L.O., et al. SMOTEBoost: Improving Prediction of the Minority Class in Boosting[A]. Cavtat-Dubrovnik, In Proceedings of Principle of Knowledge Discovery in Databases[C]. Croatia, 2003: 107-119

[10]Wang H.Y. Combination Approach of SMOTE and Biased-SVM for Imbalanced Data sets[A]. 2008 IEEE International Joint Conference on Neural Networks[C], 2008: 228-231