International Journal of Electrical and
Electronics Engineering (IJEEE)
ISSN(P): 2278-9944; ISSN(E): 2278-9952
Vol. 7, Issue 4, Jun – Jul 2018, 1-8
© IASET

International Academy of Science,
Engineering and Technology
Connecting Researchers; Nurturing Innovations
IASET

IMPROVE EFFICIENCY OF CANCER CLASSIFICATION BY COMBINING SELECTED FEATURE AND ADDITIONAL ELEMENTS

***Duong Hung Bui*[1]*, Manh Cuong Nguyen*[2]*, Thi Hong Nguyen*[3] *& Xuan Tho Dang*[4]**

[1]*Hanoi Trade Union University, Vietnam*

[2,3,4]*Faculty of Information Technology, Hanoi National University of Education, Vietnam*

ABSTRACT

In fact, many problems with imbalanced data mean that the number of elements of a class is much larger than the number of elements of the remaining classes. This is the major reason for the declining performance of data classification. In addition, we found that in a number of imbalance datasets have many features redundant, unnecessary, not important to predict. Some reports have indicated if removing these features, it will increase the accuracy in imbalance data classification. Therefore, this paper studies, data balancing methods and reduces the number of attributes to improve the efficiency of data classification. Since then, we have developed a new method to reduce the number of feature and elements in the imbalance data classification. We experimented on some sets of biological data taken from the UCI like leukemia, colon-cancer and breast-p. These results show that our new method being more accurate classifiers with the G-mean measure compared with the method using original data. In addition, we use t-test evaluation indicated a method that the results have statistical significance with the p-value on the smaller datasets 0.05.

KEYWORDS: *Cancer Classification, Features Selection, Imbalanced Data, SMOTE*

Article History

Received:* *14 Mar 2018* | *Revised:* *25 May 2018* | *Accepted:* *31 May 2018

INTRODUCTION

Classification of imbalance data is one of the difficult problems that interested by communities of machine learning and data mining. Class imbalance is usually solved on binary classification problems (only 2 classes) in which a class that we interest accounting for a very small proportion compared with the rest of the classes. In many practical applications, such as detecting fraudulent transactions [1], network intrusion detection [2], the risks in management[3], the sheet tape classification or medical diagnosis [4] [5], the class imbalance has a great influence on the efficiency of the classification models. For example in the field of network intrusion detection, the number of network intrusions is typically a very small fraction of the total number of network transactions. Or in medical databases, the classified of the pixels in the X-ray film [6] have cancer or not, the unusual pixels (cancer) accounted for only a small part of the whole image [7] [8]. This occurred on datasets of classification problem would make the classifier model learning difficulties encountered in predicting the minority class data. Most classification algorithms such as decision trees [9], SVM [10] (Support Vector Machine) [11] was designed to get overall accurate, not interested in any class. Because of this, the classifier learning algorithm for imbalance datasets encountered problems forecasting to lose minority class despite the accuracy is very high overall. For example, collection data for disease forecasting element A is 40000, in which patients class is a minority (or we interest as the positive class) with only 10 molecules and remaining classes (not disease, or negative class) has 39990 elements. A forecasting algorithm completely wrong patient A (always forecasting the disease is not A) but for the overall forecast is 99.975% [12]. This is one of the serious mistakes of classification algorithms. Because of this, solving the data imbalance classification problem is interested in a lot of scientists in the machine learning community.

Many solutions have been proposed to solve the problem above in learning algorithm to improve minority class but do not lose the majority class [13]. Many data-resampling methods were proposed to increase minority class which SMOTE [14] algorithm is one of the famous algorithms, typical and widely applied in the machine learning community. There are also proposed methods for reducing sampling majority class. Proposed changes to the data partition function improves minority class forecasting, but do not take a majority class forecasting.

In real data, in particular as bioinformatics, the appearance of the data imbalance is inevitable. Besides that we find the emergence of more and more data sets very large of attributes, although these features have a lot of redundant features is not useful in predicting the minority class. That led to the minority class prediction not good but very time consuming to run the dataset [15].

Therefore, we propose the method of selection feature and smote to help improve minority class prediction, but not losing the majority class prediction. On the other hand, does not take much time to run the data set. The experimental results on 03 imbalanced data sets in UCI repository showed that the method that we propose for greater efficiency when compared with only using SMOTE algorithm, based on the criterion of G-using mean and t-test showed that this finding is statistically significant with a p-value less than 0.05.

ATTRIBUTE SELECTION METHOD AND ADDITIONAL ELEMENTS

Measures to Evaluating Effectiveness Classification

Most research in imbalance fields are mainly focused on two-class problem, the multi-class problem can be a simplified two-class problem. Conventionally, the label of minority class is positive (positive class), and the label of the majority class label is negative (negative class).

Table 1 shows the 2-class confusion matrix [7]. TP and TN denoting the number of positive and negative samples were correctly classified, while the FN and FP denote the number of samples misclassified positive and negative respectively.

Table 1: Confusion Matrix

	Predicted as Positive	Predicted as Negative
Actually Positive	TP	FN
Actually Negative	FP	TN

Several measurements are calculated by the values in the confusion matrix [1]:

Accuracy = (TP+TN)/(TP+FN+FP+TN) (1)

FP rate = FP/(TN+FP) (2)

TN rate=TN/(TN+FP) (3)

G-mean=sqrt (TP rate*TN rate) (4)

Formula (1)): If the data is extremely imbalance, most of all element of majority class is predicted exactly and not exactly all minority samples, then the accuracy is still high because the number of majority samples is larger than the number of minority samples. In this context, the accuracy cannot reflect the predicted reliability for the minority class.

FP rate (Formula 2) represents the percentage of misclassified negative samples. TN rate (Formula 3) is the percentage of correctly classified negative samples. G-mean is determined based on two values of TP rate and TN rate (Formula 4). That is the measure of the classified efficiency of both minority and majority classes. Therefore, we use this measure to evaluate the efficiency of classification on imbalanced datasets.

Feature Selection

In machine learning and data mining, the data could contain many features that are either redundant or irrelevant which cause to reduce accuracy performance of classification. Thus feature selection techniques were proposed to remove these features without losing information. Moreover, these techniques have some advantages, for examples simple and easy to explain by researchers, shorter training times, reduce much data dimension, limit over-fitting [x, y] [16] [17].

A feature selection algorithm tries to find new feature subsets by using an evaluation measure which scores the different feature subsets. So, evaluation measure is an importance of the feature selection algorithm. Some common measures include Pearson product-moment correlation coefficient, the pointwise mutual information, the mutual information, inter/intra class distance or the scores of significance tests for each class/feature combinations.

The Pearson's correlation feature selection is the famous and effective measure evaluates subsets of features based on the hypothesis: "Good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other". A feature subset S consisting of k features is measured by the following equation [x, y] [18] [19] [20]:

$$G_S = \frac{k.\overline{r_{cf}}}{\sqrt{k + k(k-1).\overline{r_{ff}}}}$$

Figure 1: Formula for Calculating Correlation between Attribute

Here, $\overline{r_{cf}}$ is the average value of all feature-classification correlations, and $\overline{r_{ff}}$ is the average value of all feature-feature correlations. The correlation feature selection criterion is defined as the maximum G_S

SMOTE

There are many methods of re-sampling to balance the data in the classification problem such as[21]:

Under-sampling is reducing the number of elements of majority class to balance data. Random under-sampling is the simplest method. There are also several proposed advanced approaches such as the Condensed Nearest Neighbor Rule (CNN), the Neighborhood Cleaning Rule (NCL), the Edited Nearest Neighbor Rule (ENN), the To mek links [21].

Over-sampling is method increasing the number of minority elements to balance data. The simplest method is random over-sampling: randomly selects minority elements to produce identical replicas that increase minority size. However, this method increases the over-fitting of the classification model with the training dataset. Besides, there is also a way to increase the sample size by generating artificial elements and labeling them as minority class. SMOTE is the first generating artificial elements algorithm and has proven its effectiveness through experimentation.

SMOTE (Synthetic Minority Over-sampling Technique) is an over-sampling algorithm by adding synthetic elements on the juncture between the two original elements of a minority class. This approach was inspired by a technique proven successful in handwritten character recognition. In that method, they extend the training dataset by performing a rotating or tilting of the original data [14].

In this method, the authors re-sampling the minority class elements by:

- For each minority member, identify its nearest neighbor k in the minority class;
- Randomly select a neighbor in the neighboring k specified above;
- The artificial elements (synthetic elements) is generated on the line connecting the considered element and its neighbors as follows: Calculates the difference between the attribute vector of the considered element with its neighbors; Take this difference multiplied by a number between 0 and 1 (randomly selection); Then, adding this result to the attribute vector of the considered element, we obtain the attribute vector of the artificial element, which is assigned the class label attribute as the minority (Negative).

Depending on the amount of sampling required for the minority class, neighbors are randomly selected from k nearest neighbors. For example, if the number of artificial elements is twice that of the original minority class (200%), then for each of the original elements, we define two neighbors randomly in the k nearest neighbor, for each of these neighbors we make two artificial elements on the line connecting the considered element and its neighbors. Similarly, if we need to make more artificial elements, we choose more neighbors.

COMBINATION METHOD: SMOTE-CORRELATION

In order to improve the efficiency of the classification, we combine the two methods: feature selection and generate additional elements. After performing the feature selection, the dimension of the data is significantly reduced, eliminating the attributes that less affects the class prediction. Then, the new dataset was used as an input to the SMOTE algorithm to balance the amount elements between the majority and minority classes. Finally, the dataset obtained after the implementation of SMOTE will be classified by the classification algorithm as: K-NN, SVM…Figure 2 show the model represents the process of combination method.

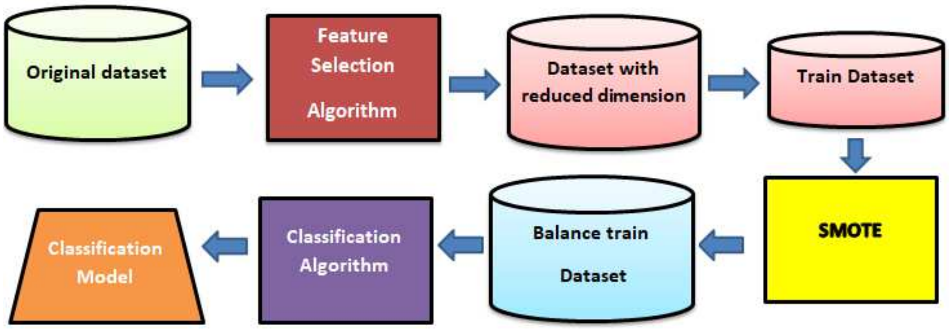

Figure 2: Model Show Combination Method: Smote-Correlation

RESULTS

To evaluate the effectiveness of the method we mentioned above, we install and run the program language R and Perl[22]. On 03 experimental imbalance cancer datasets from UCI (University of California, Irvine) and described in

Table 2. These datasets are divided into two parts: training data and test data.

- **Train Dataset:** Used for the learning process to build the classification model.
- **Test Dataset:** Used to evaluate the classification effect.

In this paper, we conducted experiments using the k-fold cross-validation method, with k = 5 (5-fold) [23]. Firstly, to reduce the number of attribute of datasets, the Correlation method was carried out with removal rates between 0.2 and 0.9, respectively. After that, we perform data balancing by the SMOTE algorithm. Finally, the classification algorithms are used K-NN, SVM using the available packages in R: Class, Kernlab [24].

Table 2: Imbalance Datasets

Datasets	Number of Elements	Number of Feature	Minority Class Ratio
Breast-p	198	32	23.73%
Colon-cancer	62	2000	35,48%
Leukemia	71	7128	34,72%

To show clearly the effectiveness of the method that we propose, we compared the results based on G-mean measure. Figure 3 is a chart denoted G-mean values of 3 original datasets (without SMOTE and Correlation) with datasets after running methods: the original data using SMOTE (SMOTE), original data using feature selection (correlation), and using a combination of data selection and additional elements SMOTE (smote-correlation). From the results in Figure 2, we can see that method applied in combination has highly effective than traditional methods. Specifically, after applying combination method, Leukemia dataset with 7128 features obtained G-mean value 87.93 and the Breast-p dataset obtained G-mean value 65.22 are higher than other methods.

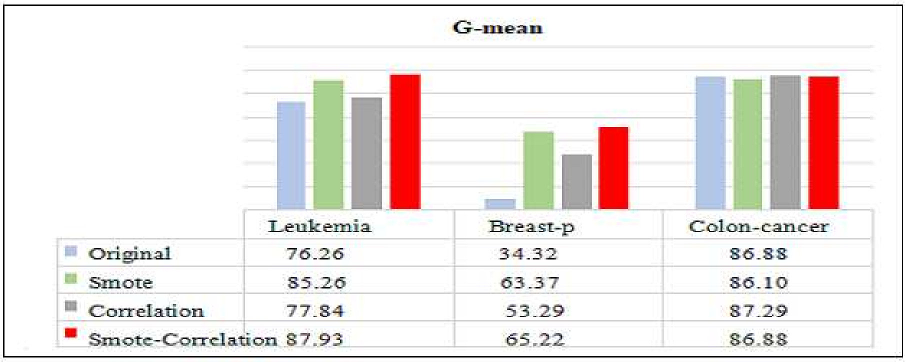

Figure 3: Chart Show G-Mean Values of 3 Dataset: Leukemia, Breast-P and Colon-Cancer

We evaluate the statistical significance of the data by using t-test to calculate the p-value for values of G-mean when we perform classification datasets after applyingSMOTE - correlation method with other methods.

The results show that almost p-values are less than 0.05 (statistically significant). Specifically, with the Leukemia dataset, the p-value of the G-mean value of the SMOTE-correlation method with Original, Correlation, SMOTE are: 2.2e-16, 2.2e-16, 3.05e-7. With the Breast-p dataset, the corresponding p-value are: 2.2e-16, 1.25e-5, 2.2e-16.

Thus, the method we propose SMOTE-Correlation has better experimental results than other methods and the results are statistically significant.

CONCLUSIONS AND DEVELOPMENT

In recent years, learning with imbalance datasets received much attention in the two aspects of theory and practice. However, the traditional data mining method unresolved imbalance data problem in a satisfactory manner. In order to solve this problem, combining feature selection method and smote in this report is a good approach to improve class efficiency.

The idea that we propose is based on leverage advantages reduce the number of a feature in the data then increase the number of elements in the minority class so that improves minority class prediction. The experimental results show that our methods are more effectively than traditional methods.

While we try to enhance technical selection features, optimization of the key feature in the imbalance data contributes to accurately assess, evaluate effectiveness in minority class. At the same time, along with computing of G-mean values of the datasets, we also compute the runtime for each dataset.

In addition, we will also experiment with the new methods and on different datasets, combination other feature selection method with SMOTE and SMOTE's improved algorithms such as: Borderline-SMOTE, Safe-level-SMOTE, Add-border-SMOTE...[25][26][27].

ACKNOWLEDGEMENT

This research was supported by the Vietnam Ministry of Education and Training, project B2018-SPH-52.

REFERENCES

1. *Mahmoudi, Nader, and E. Duman, “Detecting credit card fraud by modified Fisher discriminant analysis,” Expert Syst. with Appl. 42.5, pp. 2510–2516, 2015.*

2. *Suthaharan and Shan, “Big data classification: Problems and challenges in network intrusion prediction with machine learning,” ACM Sigmetrics Perform. Eval. Rev., pp. 70–73, 2014.*

3. *O. V. Antipina and A. C. Prokopyeva, “Classification of financial risks and management techniques in the organization of production processes,” World Sci. Discov. Mire Nauchnykh Otkrytiy, no. 65, 2015.*

4. *C. Wang and et al, “imDC: an ensemble learning method for imbalanced classification with miRNA data,” Genet. Mol. Res. 14.1, pp. 123–133, 2015.*

5. *Krawczyk, Bartosz, and et al, “Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy,” Appl. Soft Comput., vol. 38, pp. 714–726, 2014.*

6. *Lang, Philipp, and et al, “Methods and devices for evaluating and treating a bone condition on x-ray image analysis,” U.S. Pat., no. 9, pp. 275–496, 2016.*

7. *Y. Sun, A. K. C. Wong, M. Kamel, and S., “Classification of Imbalanced Data: a Review,” Int. J. Pattern Recognit. Artif. Intell., vol. 23, no. 4, pp. 687–719, 2009.*

8. *Khatami, Amin, and et al, “Parallel deep solutions for image retrieval from imbalanced medical imaging archives,” Appl. Soft Comput., no. 63, pp. 197–205, 2018.*

9. *Lior and Rokach, “Data mining with decision trees: theory and applications,” World Sci., vol. 81, 2014.*

10. Shawe-Taylor, John, and S. Sun, “A review of optimization methodologies in support vector machines,” Neurocomputing, vol. 74, no. 17, pp. 3609–3618, 2011.

11. S. B. Kotsiantis, “Supervised Machine Learning: A Review of Classification Techniques,” Informatica, vol. 31, pp. 249–268, 2007.

12. Đ. T. N. Phan Bích Chung, “Phân Lớp Dữ Liệu Không Cân Bằng Với Roughly Balanced Bagging,” Tạp chí Khoa học - Đại học Cần Thơ, pp. 189–197, 2011.

13. H. HE and E. a. Garcia, “Learning from Imbalanced Data Sets.,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263--1264, 2010.

14. N. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “{SMOTE}: {S}ynthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

15. L. C. Molina, L. Belanche, À. Nebot, J. Girona, and C. N. C, “Molina et al. - Unknown - Feature Selection Algorithms A Survey and Experimental Evaluation E if EG edG H AX with distribu.pdf,” Data Mining, 2002. ICDM 2002. Proceedings. 2002 IEEE Int. Conf., pp. 306–313, 2002.

16. Gareth James, D. Witten, T. Hastie, and R. Tibshirani, “An Introduction to Statistical Learning,” Springer, p. 204, 2013.

17. M. L. Bermingham et al., “Application of high-dimensional feature selection: evaluation for genomic prediction in man,” Sci. Rep, vol. 5: 10312., 2015.

18. M. Hall, Correlation-based Feature Selection for Machine Learning. 1999.

19. Senliol, Baris, and et al, “Fast Correlation Based Filter (FCBF) with a different search strategy,” Comput. Inf. Sci. Isc., vol. 23, no. 8, 2008.

20. H. Nguyen, K. Franke, and S. Petrovic, “Optimizing a class of feature selection measures,” in Proceedings of the NIPS 2009 Workshop on Discrete Optimization in Machine Learning: Submodularity, Sparsity & Polyhedra (DISCML), 2009, p. 5.

21. X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, “On the Class Imbalance Problem,” Int. Conf. Nat. Comput, vol. 4, pp. 192–201, 2008.

22. and B. d F. R. L.Schwartz, T. Phoenix, Learning Perl. O’reilly, 2008.

23. Arlot, Sylvain, and A. Celisse, “A survey of cross-validation procedures for model selection,” Stat. Surv., vol. 4, pp. 40–79, 2010.

24. A. Karatzoglou and et al, “Package Kernlab Version 0.9-22. An S4 Package for Kernel Methods in R. Reference Manual,” J Stat Softw, pp. 1–20, 2015.

25. H. Han, W. Wang, and B. Mao, “Borderline-SMOTE : A New Over-Sampling Method in,” ICIC, pp. 878–887, 2005.

26. C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, “Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem,” Lect. Notes Comput. Sci. (including

Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5476 LNAI, pp. 475–482, 2009.

27. *H. NT, C. NM, and T. DX, "Add-border-smote: phương pháp mới sinh thêm phần tử trong phân lớp dữ liệu mất cân bằng," Tạp chí khoa học và kĩ thuật - Học viện kĩ thuật quân sự, vol. 164, no. 1, pp. 81–91, 2014.*