Volume 3, Issue 11, November 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Data Preprocessing for Liver Dataset Using SMOTE

K.Lokanayaki
Assistant Professor
Florence Group of Institutions Bangalore, India

Dr.A.Malathi
Assistant Professor
Govt. Arts College Coimbatore, India

Abstract:-The class imbalanced problem occurs in various disciplines when one of target classes has a small number of instances compare to other classes. A classifier normally ignores or neglects to detect a minority class due to the small number of class instances. It poses a challenge to any classifier as it becomes hard to learn the minority class samples. Most of the oversampling methods may generate the wrong synthetic minority samples in some scenarios and make learning tasks harder. To overcome this problem in the minority samples first identify the missing attribute data in correctly and learning the task easier. In this paper purpose a new setting of missing data imputation and achieve a high classification rate of Imbalanced liver datasets. To achieve a high classification rate using evolutionary based oversampling, undersampling, Synthetic Minority Over-sampling Technique results are applied to classification using SVM.

Keywords: Imbalanced Dataset, SMOT, Undersampling, Oversampling, SVM

I. INTRODUCTION

The imbalanced dataset problem in classification domains occurs when the number of instances that represents one class larger than the other class. It has two types of problems. First we cannot analyze default inability of common classifiers in minority class and majority class. Second have different cost of misclassification. So researches found number of standard classification algorithms for imbalanced dataset. Mostly overcome these type of problems can be used sampling approach [1] [11]. Sampling is one of the basic approach. We can choose to alter imbalanced dataset, we can choose sampling. Because sampling can alter data in imbalanced dataset. It have two types of sampling contains under sampling and oversampling. Under sampling used for removing instances in set of majority class. Oversampling used for add the instances of minority class. like SMOTE [2], which are able to create new synthetic examples fitting to the minority class, and widen the decision region for the classifier.It also [3] creates random synthetic minority instances along the line segments connecting a minority instance and its neighbors. It is claimed that SMOTE can generate more general decision regions for the minority class.

II. RELATED WORK:

A new method in imbalanced dataset for preprocessing through the construction of a Synthetic Minority Oversampling Technique (SMOTE) with Rough set theory proposed by Martin Hlosta at al [1]. Finally they are combining this technique and c4.5 method produced good result analysis. Similar to [2] SMOTE based oversampling and evolutionary undersampling technique with c4.5 and PART method for imbalanced dataset. Alberto Fernández at al [3] also used in safe level SMOTE for produced good accurate result. In [4] analyzed SMOTE with c4.5, Ripper and Naïve Bayes classifier for better performance. Jai li at al found classification performance using Random – SMOTE (R-S). This method to increase number of the random minority samples [5][13]. However this type of learning the class imbalanced problem with several oversampling sampling and undersampling methods, before that the missing attribute values are found and the best minority class features are selected to classify imbalance data. To find the missing values of each instance are imputed by considering the number of instances that are most similar to the instance of interest .KNN is one of the method [9] to find the missing values of an instance are imputed considering a given number of instances that are most similar to the instance of interest. The similarity of two instances is determined using a distance function. However the followings are main drawback,

- To deal with missing values in data sets with homogenous attributes KNN methods are independent of all either continuous or discrete value.
- Even though the existing systems are presented for the imputation of the missing value attribute, they have several drawbacks.
- In the research of the missing value imputation, the existing systems are not well defined for the mixed attributes.

III. FRAMEWORK OF PROPOSED WORK

The proposed framework is shown in figure1. The information is obtained from UCI Machine Learning Repository [15], all records contains in database in which the data may be redundant, noisy or irrelevant in nature. The proposed pre-processing approach filters data effectively and the result compares with existing approach.

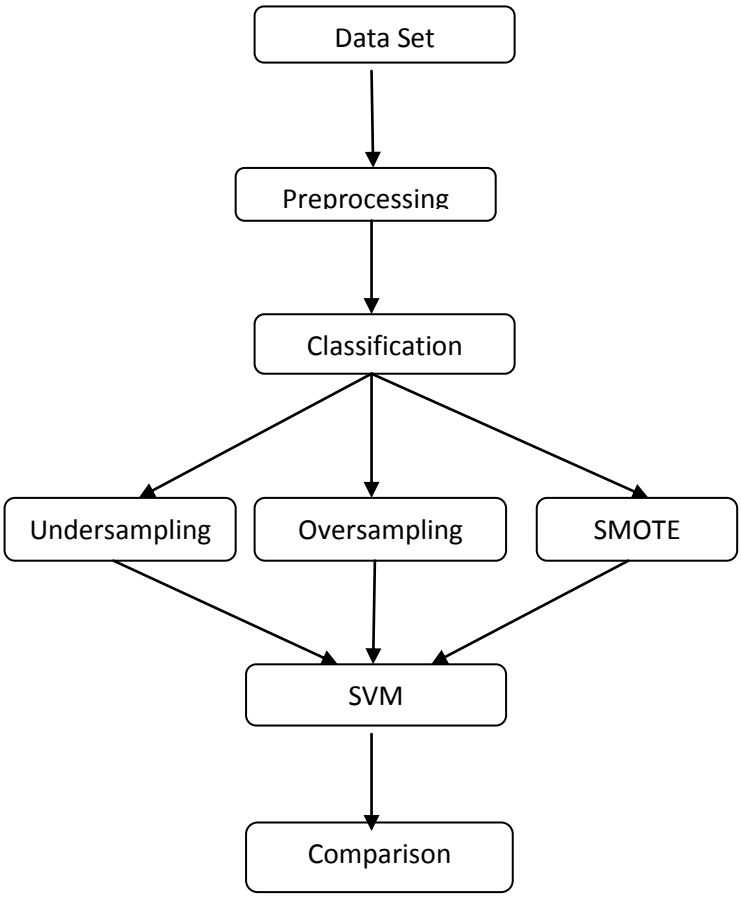

Fig. 1 Framework for proposed work

IV. PROPOSED SYSTEM

Missing data imputation is a key issue in learning from incomplete data. Existing KNN technique have been developed to deal with missing values in data sets with homogenous attributes. But this approach is independent of all either continuous or discrete value. Proposed a new setting of missing data imputation that is by imputing missing data in data sets with heterogeneous attributes thus by contributing both continuous and discrete data. Here they propose two consistent estimators for discrete and continuous missing target values. And then, a mixture kernel based iterative [10] estimator is advocated to impute mixed-attribute data sets. In this a kernel functions for the discrete attributes are studied and then a mixture kernel function is proposed by combining a discrete kernel function with a continuous one. In this method the input dataset is considered as the liver cancer data to perform the imbalance dataset learning process .First identify the missing values attributes using the data imputation it is denoted as MVi and the imputed value of MVi in nth iteration imputation is regarded as (Mv)ˆ_i^t.

From the above algorithm, all the imputed values are used to impute subsequent missing values, i.e., the (n+1) iteration imputation is carried out based on the imputed results of the nth imputation, until the filled-in values converge or begin to cycle or satisfy the demands of the users.

//the first iteration

1.Missing attribute imputation in the dataset

1.1 For each missing value attributes imputation (MVi) in the both discrete (Y) and continuous case(Y)

M̂V_i^1=mode(S^r in Y) // If Y is discrete variable

M̂V_i^1=mean(S^r in Y)

//if Y is continuous variable

End for

2.Perform n iteration for imputation(n>1)

2.1 Initially n=1

2.2 REPEAT

2.3 n++;

2.4 for each MVi in Y

〖MV〗_i=M̂V_i^(t-1), p ∈ S_m ,p=1,…m ,p≠i

2.4.1 The missing value attribute of the nth imputation is evaluated based on the equation

Yˆ_i^n=mˆ_t (X_i)+ϵ_i^n

mˆ_n (X_i) is the kernel estimator for m_n (x)(xϵ R^(d+p)) based on the complete pairs (X^n,Y^n) and ϵ_i^t is simple random size with m with replacement {Y_i^n-mˆ_t (X_i) }i∈S_r //discrete variable

End for
Until
|〖CA〗_n-〖CA〗_(n-1) |≥∈
Convergence or cycling
Output
n//n number of iterations
completed dataset

Fuzzy rule based classification systems [6] using a preprocessing step in order to deal with the class imbalance. Our aim is to analyze the behavior of fuzzy rule based classification systems in the framework of imbalanced data-sets by means of the application of an adaptive inference system [7] with parametric conjunction operators.Any classification problem consists of m training patterns xp ={xp1; ... ; xpn}, p = 1; 2; ... ;m from M classes where xpi is the ith attribute value (i = 1; 2; ... ; n) of the pth training pattern. In this work, fuzzy rules of the following form for our FRBCSs:

Rule Rj : If x1 is Aj1 and ... and xn is Ajn then Class = Cj with RWj

Where Rj is the label of the jth rule, x =(x1; ... ; xn) is an n-dimensional pattern vector, Cj is a class label, Aji is an antecedent fuzzy set and RWj is the rule weight. Then use triangular membership functions as antecedent fuzzy sets[12][8].

The following steps are feature space of the classes:

(1) Establishment of the linguistic partitions. Once the domain of variation of each feature Ai is determined, the fuzzy partitions are computed.

(2) Generation of a fuzzy rule for each example xp =(xp1; ... ; xpn); Cp. To do this is necessary

2.1) To compute the matching degree μ(x_p)of the example to the different fuzzy regions using a conjunction operator (usually modeled with a minimum or product t-norm).

(2.2) To assign the example xp to the fuzzy region with the greatest membership degree.

(2.3) To generate a rule for the example, whose antecedent is determined by the selected fuzzy region and whose consequent is the label of class of the example.

(2.4) To compute the rule weight.

Sampling is one of the traditional methods to balance the data set. Evolutionary Random under Sampling, Evolutionary Random Over sampling, SMOTE sampling is used to overcome the equitable problem. SVM classifier [14] is great deal of success in a variety of areas such object recognition form classification, liver cancer dataset. SVM classifier is used to build a model based training data. The result of training data and testing data is measured at each classification dataset.

V. IMPLEMENTATION

The analyses done on UCI Machine Learning Repository for liver disease training dataset and the accuracy of the proposed method is compared with Evolutionary Random under Sampling, Evolutionary Random Over sampling. In this paper predicted accuracy is 96.2% in 5th iteration showed in figure.2 .Existing techniques predicted accuracy 94%, 91.4%, in 5th iteration.

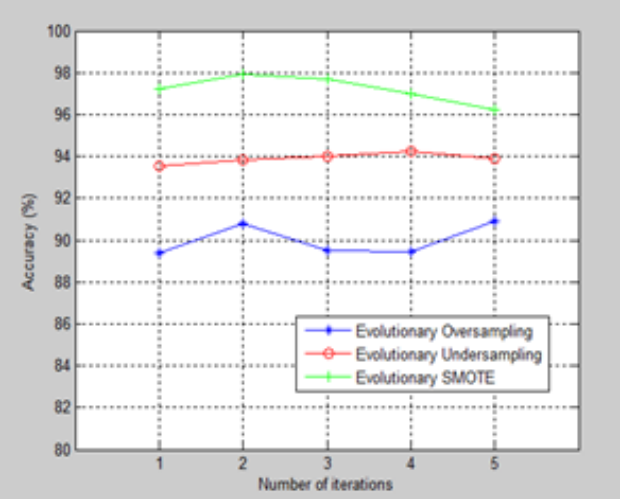

Fig.2. accuracy of proposed technique

VI. CONCLUSION

This paper analyzed the performance of liver dataset and it proposed a hybrid SMOT technique to overcome the imbalanced dataset difficulties. In this paper proposed compare evolutionary based oversampling, undersampling, and hybrid Synthetic Minority Over-sampling Technique results are applied to SVM for a high classification rate. Finally this paper proposed SMOT technique will produce high classification rate compare than oversampling and undersampling.

REFERENCES

[1] Martin Hlosta, Rostislav Stríž, Jan Kupčík, Jaroslav Zendulka, and Tomáš Hruška "Constrained Classification of Large Imbalanced Data by Logistic Regression and Genetic Algorithm" *International* Journal of Machine Learning and Computing, Vol. 3, No. 2, April 2013.

[2] V. Chawla ,W. Bowyer Nitesh,Lawrence O. HallKevin, W. Philip Kegelmeyer "Synthetic Minority Over-sampling Technique" *Journal of Artificial Intelligence Research* 16 (2002) 321–357.

[3] Alberto Fernández , María José del Jesus, Francisco Herrera "On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets", *Expert Systems with Applications* 36 (2009) 9805–9812.

[4] Julia´n Luengo, Alberto Ferna´ndez ,Salvador Garcı´, Francisco Herrera ,"Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary undersampling", *Applied Soft Computing*(2011) 15:1909–1936.

[5] Enislay Ramentol ,Yailé Caballero, Rafael Bello, Francisco Herrera," SMOTE-RSB a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory", *Springer-Verlag London Limited* 2011

[6] Chin-Yuan Fana, Pei-Chann Changb, Jyun-Jie Linb, J.C. Hsiehb,"A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification", *Applied Soft Computing* 11 (2011) 632–644.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer,"SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16(1), pp. 321—357, 2002.

[8] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special Issue on Learning from Imbalanced Data Sets," *ACM SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets*, vol.6(1), 1—6, 2004.

[9] Fernández, A., García, S., del Jesus, M. J., & Herrera, F," A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets", *Fuzzy Sets and Systems*, 159(18), 2378–2398,2008.

[10] Ghosh, A., Pal, N., & Das, J," A fuzzy rule based approach to cloud cover estimation. Remote Sensing of Environment", 100(4), 531–549,2006.

[11] Wang, L. X., & Mendel, J. M.." Generating fuzzy rules by learning *", IEEE Transactions on Systems, Man, and Cybernetics*, 25(2), 353–361,1992

[12] Orriols-Puig, A., & Bernadó-Mansilla, E,"Evolutionary rule-based systems for imbalanced datasets", *Applied Soft Computing*, 13(3), 213–225,2009.

[13] Ishibuchi, H., & Yamamoto, T," Rule weight specification in fuzzy rule-based classification systems", IEEE *Transactions on Fuzzy Systems*, 13, 428–435,2005.

[14] ZhiQiang Zeng,Ji Gao,"Improving SVM Classification with Imbalance Data Set*" Proceedings of the 16th International Conference on Neural Information Processing*" Pages 389-398 ,2009

[15] Blake, C., Merz, C.: UCI Repository of Machine Learning Databases. Department of Information and Computer Sciences, University of California, Irvine, CA, USA (1998),

