ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

AN IMPUTE MISSING VALUES USING IMPROVED WEIGHTED SMOTE FOR LIVER CELL IMBALANCED DATASET

K.Lokanayaki

ABSTRACT

In research field, the main problem is class imbalanced problem and impute missing attribute. In an imbalanced dataset occurs in various restraints when one of target classes has a small number of instances compare to other classes. Most of the oversampling methods may generate the wrong synthetic minority samples in some scenarios. To overcome this problem in the minority samples first identify the missing attribute data in correctly and learning the task easier. In this paper proposes the extension of Weighted SMOTE called an Improved Weighted Synthetic Minority Oversampling Technique (IWSMOTE), which can overcome the problem of finding the missing attribute value of each samples for imbalanced liver cell dataset. The proposed algorithm evaluated based on experimental study. This algorithm compared against a existing SMOTE and Weighted SMOTE generalizations.

KEYWORDS: Imbalanced Dataset, SMOTE, Weighted SMOTE, Oversampling.

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 difficulty invents from the absence of some group oftest in the database. This problem, which occurs mainly for patients, is due either to a lack of time (time constraint of the clinical routine) or because the patient's inability to perform the test. Therefore, some samples in the dataset can have a high percentage of missing data. To handle this problem using sampling approach [1] [11].

Sampling is one of the basic approaches. 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. Undersampling 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.

RELATED WORK

At present, the most commonly used for imputation methods for dealing with Missing Attribute Values (MAV), each MAV is replaced by a value generated by non missing values. Many research scholars have proposed a variety of MAV imputation methods. In [11] *K*-nearest neighbours (K-NN) imputation uses to impute the MVs. In this method every time finds a MV, computes the *K*-nearest neighbours of the none missing values and imputes a value from them. It's also taken most common value among all neighbours and for numerical values using average value. RI method [12] selects some self-sufficient attributes for predicting the MV.

They found first regression equation to estimate the MV then replaces the MV value. But EM [10] method finds the maximum likelihood estimates according to the observed data was consists of following steps: expectation step (*E*-step) and maximization step (*M*-step). In first step calculates the expectation of the complete sufficient data, current parameter estimates and updates the parameter estimates. In second step repeats the two steps till the parameter estimation and the expectation of each MV imputation.

The [13] Rubin has proposed multiple imputation (MI) method. Krause and Polikar [14] have proposed e*nsemble based method for missing attribute values.* Mohammed et al. [15] have also proposed Learn++ method for missing features (abbreviated as LMF). In this method selects a number of attribute subsets from whole attribute sets, for each instance, this method finds better results of the selected attributes. Many researches impute missing attribute values for imbalanced dataset. In [8] authors have proposed CART algorithm. This algorithm handles only numerous values. Author [9] proposed missing imputation for samples to predictions with a high percentage of missing data,

Synthetic Minority Oversampling Technique (SMOTE)
It was developed by Chawla, Hall, & Kegelmeyer in 2002 [2], is an over-sampling technique. In this technique randomly generate synthetic minority examples. It also combines conversant over-sampling of the positive class with random under-sampling of the majority class. Using the over-sampling approach the minority class is over-sampled by creating artificial examples of *k* nearest class neighbors. In this technique, take the difference between the each sample under consideration and its nearest neighbor. Multiply this difference by a random number between 0 and 1. In this process for a selection of a random value with a two specific features for generating each synthetic sample.

Martin Hlosta at al [1] proposed a new method for imbalanced dataset of preprocessing through the Synthetic Minority Oversampling Technique (SMOTE) with Rough set theory. Finally they are combining proposed technique with c4.5 method produced good result for classification. 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 for classification. 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, the SMOTE has some problem. It selected value based on probability value for entire K samples from minority and majority class. Proposed Weighted SMOTE for finding missing attribute values. This technique assigns a weighted value for each class, when generated a new samples. Weighted value calculated by following formula:

$$w_i = \left(\frac{max_{j=1}^{k}(n_j)}{n_i}\right) / \sum_{j=0}^{k} n_j \quad i,j = 1,2 \ldots\ldots K \quad \text{---------- (1)}$$

It is used to normalize the weighted value, stored in the range of 0 and 1. In this formula weight value calculated only majority class.

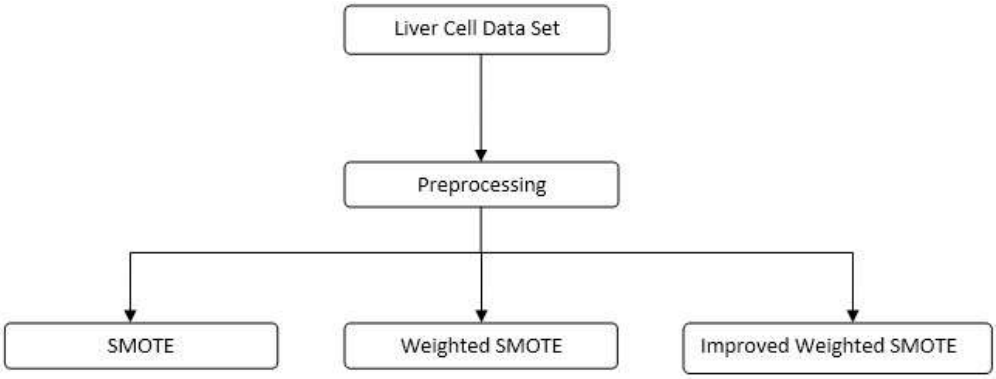
Even though SMOTE and hybrid SMOTE achieves a better results of the number of samples in each class for pre-processing and classification. Similarly, existing methods for finding the missing values of each instance is imputed by considering the number of instances that are most similar to the instance of interest. SMOTE is one of the method [9] to find the missing values of an instance are imputed in each class, when used in separation it may obtain results that are not as good as 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,

- Most of existing methods for finding missing values in data sets with homogenous attributes. Some Existing 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.
- SMOTE presents several drawbacks related to its blind oversampling, where by the creation of new positive (minority) examples only takes into account the closeness among positive examples and the number of examples of each class.

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.In this paper impute missing value calculated by weighted value. It also apply impute missing value for classification accuracy, especially class imbalance problem. We design and implementation based on oversampling technique for handling imbalanced liver cell datasets

FRAMEWORK OF PROPOSED WORK

The proposed framework is shown in figure1. These information is obtained for liver cell datasets, 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.



PROPOSED SYSTEM

Missing data imputation is a key issue in learning from incomplete data. Existing SMOTE 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)
MV̂_i^1=mode(S^r in Y) // If Y is discrete variable
MV̂_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=MV̂_i^(t-1), p ∈ S_m ,p=1,…m ,p≠i
.4.1 The
.4.2]\\
.4.3] 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

$$w_i = \left(\frac{min_{i=1}^{k}(n_i)+max_{j=1}^{k}(n_j)}{n_i}\right) / \sum_{j=0}^{k} n_j \quad i,j = 1,2 \ldots . K \quad \text{---------- (2)}$$

In this formula calculated both class weights and hold fewer samples of each class.

IMPLEMENTATION

The liver cell dataset analyses done based on proposed method accuracy for impute missing attribute values is compared with existing SMOTE and Weighted SMOTE. Proposed technique predicted accuracy is 99.23% in 50 % missing attribute values showed in figure.1 .Existing techniques predicted accuracy 96.45% and 95.41% in 50% missing attribute value.

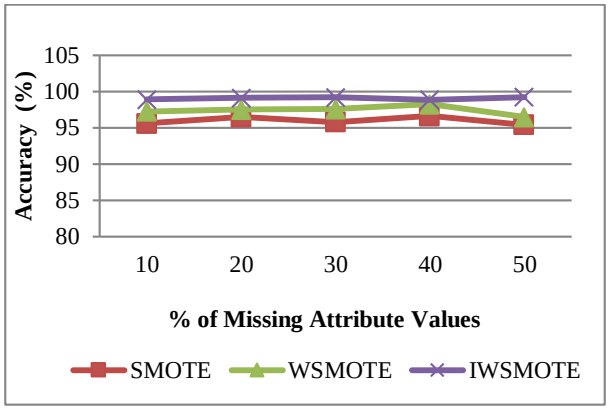


figure.1. Accuracy of proposed algorithm

As shown in the figure1 accuracy of proposed and existing algorithm based on mean value performed well in terms of 50 % missing attribute values accuracy.

CONCLUSION

In this paper analyses for an imputing missing attribute value and an imbalanced problem of liver cell dataset using Improved Weighted SMOTE algorithm (IWSMOTE). In this algorithm proposed based on sampling approach for classify the detection of liver cancer cell according to the character of the algorithms work very well. It is also used to reduce the number of the instances in the imbalanced dataset. It also reached more accuracy the classifiers for resolve missing attribute data and imbalanced data.

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] 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.

[4] 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
[5] 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.
[6] 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.
[7] Orriols-Puig, A., & Bernadó-Mansilla, E,"Evolutionary rule-based systems for imbalanced datasets", *Applied Soft Computing*, 13(3), 213–225,2009.
[8] Blake, C., Merz, C.: UCI Repository of Machine Learning Databases. Department of Information and Computer Sciences, University of California, Irvine, CA, USA (1998),
[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from in complete data via the EM algorithm," *Journal of the Royal Statistical Society B*, vol. 39, no. 1, pp. 1–38, 1977.
[10] G. E. A. P. A. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Applied Artificial Intelligence*, vol. 17, no. 5-6, pp. 519–533, 2003.
[11] J. W. Grzymala-Busse and L. K. Goodwin, "Handling missing attribute values in preterm birth data sets," in *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, D. Slezak, J. Yao, J.F. Peters, W. Ziarko, and X. Hu, Eds., vol. 3642 of *Lecture Notes in Computer Science*, pp. 342–351, 2005.
[12] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, JohnWiley & Sons, New York, NY, USA, 1987.
[13] D. B. Rubin, "Formalizing subjective notions about the effect of nonrespondents in sample surveys," *Journal of the American Statistical Association*, vol. 72, no. 359, pp. 538–543, 1977.
[14] S. Krause and R. Polikar, "An ensemble of classifiers approach for the missing feature problem," in *Proceedings of the International Joint Conference on Neural Networks*, pp. 553–556, Portland, Ore, USA, July 2003.
[15] H. S.Mohammed, N. Stepenosky, and R. Polikar, "An ensemble technique to handle missing data from sensors," in *Proceedings of the IEEE Sensors Applications Symposium*, pp. 101–105, Houston, Tex, USA, February 2006.