Multiclass Data Imbalance Oversampling Techniques (Mudiot) and Random Selection of Features

V.Shobana,K.Nandhini

Abstract: ***Class imbalance is a serious issue in classification problem. If a class is unevenly distributed the classification algorithm unable to classify the response variable, which will result in inaccuracy. The technique Multiclass Data Imbalance Oversampling Techniques (MuDIOT) is to find out the factors which have a hidden negative impact on classification. To alleviate the negative impact the technique MuDIOT concentrates on balancing the data and the result minimizes the problems raised due to uneven distribution of classes. The dataset chosen has a multiclass distribution problem and it is handled to produce better results of classification.***

Keywords: Imbalanced data, data preprocessing, big data, MuDIOT, SMOTE, RFE, random forest

I. INTRODUCTION

Classification coined as most common factor in machine learning. It is referred as the process of classifying an unpredicted, unordered value.[1]. It contains algorithms applied to the data and build the model that can be classified and discovers the dependencies behind class attributes. After that, new labels are tested and classified to the predicted groups. The data produced so far has its unique characteristics. Based on the characteristics alone we cannot build the perfect model. Most of the real world applications, particularly in healthcare, retrieving and calculating the required parameters is expensive and may not be done at all. Gathering samples from each of the above mentioned classes is difficult because of the above factors. When one class samples is more in number than that of other class, it is known as class imbalance problem. It is a common thing in medical databases when a large number of patient data is taken into consideration.[1]Multiclass Data Imbalance Oversampling Techniques (MuDIOT) deals with this problem by random sampling of minority classes. It deals with multiple classes wherein a particular class or group of classes lies under minority values. Health care data analytics requires much importance since a large amount of data is generated. There are many analysis carried out in the medical domain to improve the classification techniques and results are analyzed to increase predictor accuracy [4]–[6], especially when it is the case of uneven distribution of datasets. Thus machine learning algorithms have been applied to dataset that are not evenly distributed or incomplete, are discussed in this paper.

Revised Manuscript Received on September 05, 2019.

V.Shobana , Research Scholar, Department of Computer Science, Chikkanna Government Arts College, Tirupur, India.Assistant Professor, Department of Computer Science, Dr. N. G. P. Arts and Science College, Coimbatore, India.

Dr,K.Nandhini' Assistant Professor, Department of Computer Science, Chikkanna Government Arts College, Tirupur, India.
email: shobana484@gmail.com, krishnandhini@yahoo.com

The classes which are not evenly distributed is the major issue which is in attention until 1990s [7]. In the year 2005 dealing with imbalanced cost effective data was a major issue and it occupies the top ten 10 challenging issues in data mining.[8] Another problem in medical dataset is which attributes to choose for classification. It results in Feature Selection and a variety of methods available for selecting the top features.

II. BACKGROUND

2.1 Imbalanced Data Identification

The imbalanced class distribution defined as the relation between more numbers in majority class than that of minority class.[9].This lack of equality occurs in most of the medical databases, where different patients are diagnosed for different illness. These types of patients require special treatment. In specific cases, the datasets are fairly imbalanced with a imbalance ratio of, 1:10000 [7]. The classification and prediction algorithm shows improper classifiers and predictors on applying imbalanced dataset. The figure 1 shows the uneven distribution of classes.

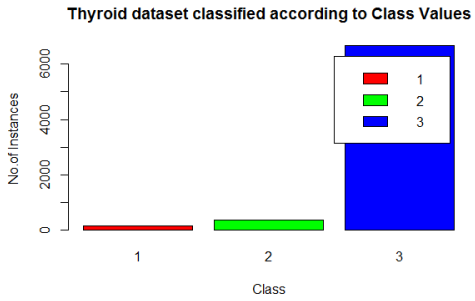


Fig.1. Imbalanced Data.

The uneven distribution of classes is given in terms of a table and is shown in Table.1.

Table.1. Imbalanced Instance Values

Class	1	2	3
No.Of Instances	166	368	6665

The main idea to resolve data imbalance is to resample the data as many times to obtain better class and even class distribution. Once it is evenly scattered it becomes easy for the classifiers to work on standard conditions. It will improve the classifier performance rapidly.

2.2 Competency Feature Selection- Removing of redundant features

Data contains attributes which are correlated with each other. Most of the methods shows better results if the correlated attributes are removed. A correlation may be positive where both variables travels in same direction or it be negative, where one attribute value increases the others decreases. Correlation may be neural or zero, which means the variables are not related. The performance of some of the algorithms become very worse if two or more variables are tightly related, called multicollinearity. An example of the above case is linear regression where the highly depended variables are removed, in order to improve the model. The proposed work after applying RFE, the correlation matrix for the thyroid dataset is shown in Table.2.

Table.2. Correlation matrix

	TSH	T3	TT4	T4U	FTI
TSH	1	-0.16	-0.26	0.07	-0.28
T3	-0.16	1	0.48	0.29	0.35
TT4	-0.26	0.48	1	0.39	0.79
T4U	0.07	0.29	0.39	1	-0.21
FTI	-0.28	0.35	0.79	-0.21	1

2.3 Ranking the Important Features

The significance of the attributes can be calculated from the data by building the model. Some algorithms such as decision trees have some mechanism to predict the important attributes that is to taken for analysis. For some other algorithms, the importance of attributes can be obtained by ROC curve analysis which is carried out on each attribute [16][10].The variable chosen for analysis is a notable output of the random forest algorithm. For every attribute in the matrix the importance of each attribute ia analyzed and the it is taken for analysis. The importance plot function gives the feature plotted on y-axis, and the variable importance plotted on x-axis. They are arranged from top to bottom as most important to least important. Hence, the most important variables are on the top and an estimate which shows their importance is mentioned by the dot position on x-axis. [18] Obviously, take a large gap across variables and choose the best one to work with. But the variables or features should be equally distributed to avoid neither over distribution nor under distribution.The example below loads thyroid dataset and constructs a Learning Vector Quantization (LVQ) model. The varImp is the measure which gives the priority of top ten attributes among 22 attributes of the dataset. These top ten attributes are used to evaluate the importance of variables which is calculated and is shown in Figure.2. It shows that the TSH, FTI and On_thyroxine attributes are the top priority attributes in the dataset and the sex attribute is the least priority attribute.

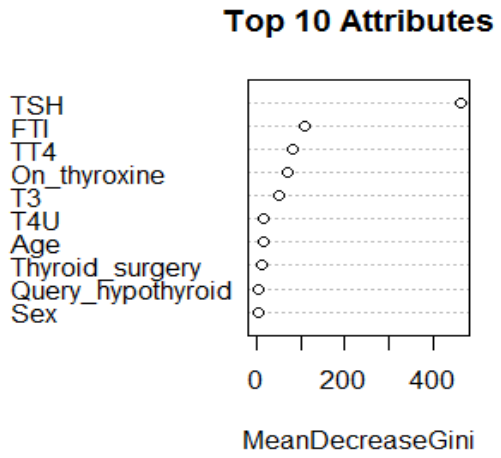


Fig.2. Variable Importance.

Gini Impurity is a measure of an incorrect classification of the random variable and if that new instance is distributed randomly which is based on distribution of class labels. The formula which calculates the Gini impurity for the randomly selected features is given in formula 1.

$$G(k) = \sum_{i=1}^{N} P(i) * (1 - P(i)) \qquad \text{--------(1)}$$

Where P (i) is the probability of classification I.

2.4 Feature Selection

The automatic selection of features can be applied to construct samples with different partitions of dataset and identifying which attributes are needed or not necessary for construction of accurate models. A widely known automatic method for selection of features is available in R caret package and is known as Recursive Feature Elimination or RFE. The illustration described below is an example of RFE method applied on thyroid dataset. A Random Forest algorithm is applied at each step to analyze the model. The algorithm is trained and tested against the combination of values of attributes. Of all the eight attributes chosen which are of different attribute sizes, only five attributes gives the comparable results [12][16].

- Recursive Selection of Features.
- Method of Outer resampling.
- Revamping of performance over a subset of attributes.

Table.3. Recursive Feature Selection

Variables	RMSE	Rsquared	MAE	RMSESD
1	0.2202	0.6271	0.06168	0.0559
2	0.1643	0.7942	0.0366	0.023
3	0.1415	0.8461	0.03319	0.0233
4	0.1391	0.8546	0.03643	0.0173
5	0.136	0.8631	0.03906	0.0156

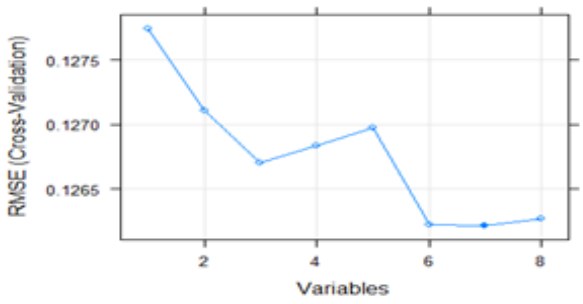


Fig.4. Variable importance

The top 5 variables (out of 5) identified out of RFE is: T4U, FTI, T3, TSH, TT4.It is plotted as graph.

III. PROPOSED METHODOLOGY- IMBALANCED DATA HANDLING

3.1 Synthetic Minority Oversampling Technique (SMOTE):

The data which are imbalanced is a supervised learning problem in which one class has more numbers than the other classes. The problem occurs frequently in binary as well as multi-level classification problems. In other words, a data set which shows distribution between the classes as unequal are considered to be imbalanced. The methods given below are used to handle imbalanced data.

- Under sampling
- Oversampling
- Synthetic Generation of Data
- Cost Effective Learning

The work depicted in this paper focuses on production of data using SMOTE. It works as creating artificial data rather than recurring and manifolding data from minority classes. It is also a method of oversampling technique. In synthetic generation of data, synthetic minority oversampling technique (SMOTE) is most common used method. The algorithm generates unnatural data based upon the feature space. It generates a group of minority class that are selected randomly and shifts the classifier learning upon minority class [14]. To generate synthetic data, it uses bootstrapping and k-nearest neighbor algorithms. It works as shown below :

- Calculate dissimilarity between the sample space and the nearest neighbor of it.
- The dissimilarity obtained is replicated and applied on number that occurs with the range 0 and 1.
- The result obtained is joined to the sample space taken for consideration..
- The random selection point is calculated across two specific features along the line segment.

3.2 Multiclass Data Imbalance Oversampling Techniques (MuDIOT).

A replacement of over and under samples known as Synthetic Minority Oversampling Technique (SMOTE) was introduced in 2002 by N.Chawla et al. [15] generates raw data. New data produced along the positive class and any of the k-nearest neighbors. SMOTE constructs a combination of oversampling the lowest class and under sampling the highest class. It proves a good efficiency in most of the cases but in rare case when data is distributed in complex. [16][10].

The binary class problem is rectified using MuDIOT.The advantage of this approach is that while standard boosting gives equal weights to all misclassified data, MuDIOT gives more examples of the minority class at each boosting step. The improved (proposed) method follows k-nearest neighbor for each sample and it vary in results.

Pseudo code of MuDIOT

Input Parameter: Dataset D, D= {x1, x2, x3, … , xn}
Output Parameter: Class = [1, 2, 3]
for each instance (x) in D Find the k-nearest neighbors of x
randomly choose one of the k-nearest neighbors
calculate the distance between the k-nearest neighbor and x.
randomly adjust the magnitude of the difference by multiplying each dimension by a random number between 0 and 1.
Add the new difference to the original instance.
Append the new instance to the synthetic Class Vector.
Repeat the above steps with different k values.

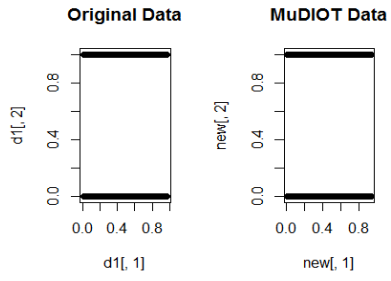
Formula to generate synthetic data using MuDIOT can be written as,

$$D_{new} = D_x + (D_k - D_x) * R \quad \text{------(2)}$$

where D_{new} denotes the MuDIOT data, D_x denotes the original dataset D_k denotes the k nearest neighbor and R denotes a random number between 0 and 1. The results for different k values are shown below. Each step improves the oversampled data than that of original instances.

IV. RESULTS AND DISCUSSION

By applying the above algorithm the dataset has been adjusted for oversampling and under sampling. By doing so the result has an ad versant effect. The data is distributed evenly on the three class attributes which will not affect the performance and accuracy of algorithms. The results are visualized as given below.



At k=1

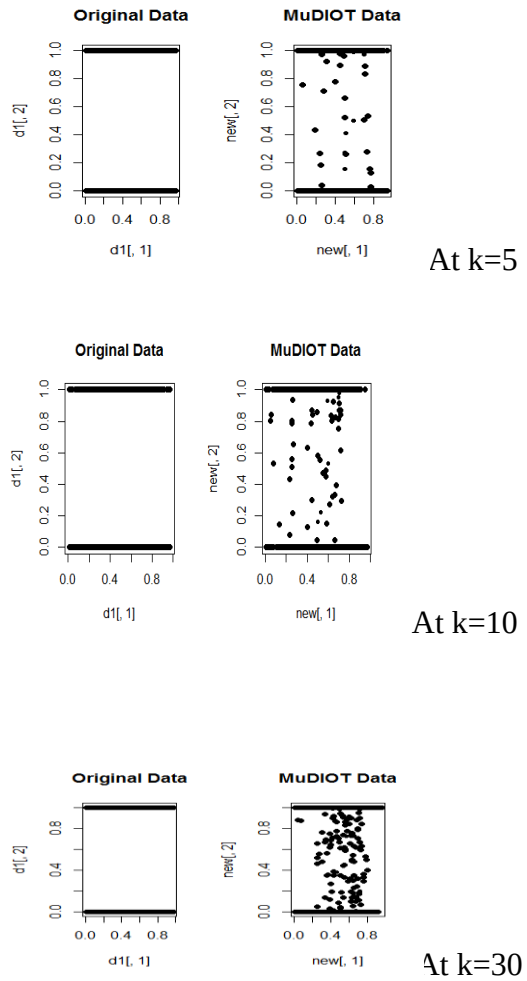


Fig.4. MuDIOT for different k values

Where k is a number which indicates the number of nearest neighbors which is used produce new examples of the minority classes. For each value of k the dataset taken under samples the imbalanced data so that it will improve the accuracy of the classification.

V. CONCLUSION

The technique discussed above overcomes the problem of uneven distribution of classes in the dataset taken. Class imbalance is a severe issue in most of the medical datasets and if it is not handled properly it leads to major issues. The dataset taken has suffered severely from class imbalance where a large amount of response variable belongs to a particular class. It is reduced using the above technique where the minority classes are being oversampled to reach the majority class. The results shows an improvement in preprocessing of data and when applied to an algorithm will produce better accuracy.

REFERENCES

1. Stefanowski J.: "Dealing with Data Difficulty Factors while Learning from Imbalanced Data", Challenges in Computational Statistics and Data Mining, 2016, pp. 333–363, DOI: 10.1007/978-3-319-18781-5_17.
2. Senthilkumar D., Paulraj S.: "Diabetes Disease Diagnosis Using Multivariate Adaptive Regression Splines", International Journal of Engineering and Technology, vol.5(5), 2013, pp. 3922-3929.
3. Arslan A.K., Colaka C.: "Different medical data mining approaches based prediction of ischemic stroke", Computer Methods and Programs in Biomedicine, 2016, vol. 130, pp. 87–92, DOI: 10.1016/j.cmpb.2016.03.022.
4. Wosiak A., Dziomdziora A.: "Feature Selection and Classification Pairwise Combinations for High-dimensional Tumour Biomedical Datasets", SchedaeInformaticae, 2015, vol. 24, pp. 53-62, DOI: 10.4467/20838476SI.15.005.3027.
5. Glinka K., Wosiak A., Zakrzewska D.: "Improving Children Diagnostics by Efficient Multi-label Classification Method", Information Technologies in Medicine 2016 vol. 1, series: Advances in Intelligent Systems and Computing 471(1), eds.: EwaPietka, PawelBadura, JacekKawa, WojciechWieclawek, Springer International Publishing, pp. 253-266, DOI: 10.1007/978-3-319-39796-2.
6. Levashenko V., Zaitseva E.: "Fuzzy Decision Trees in medical decision Making Support System" 2012 Federated Conference on Computer Science and Information Systems (FedCSIS), Wroclaw, 2012, pp. 213219.
7. He H., Garcia E. A.: "Learning from Imbalanced Data", IEEE Transactions on Knowledge and Data Engineering, 2009, vol. 21(8), pp. 1263– 1284, DOI: 10.1109/TKDE.2008.239.
8. [8]Yang Q., Wu X.: "Challenging problems in data mining research", International Journal of Information Technology and Decision Making, 2006, vol. 5(4), 597–604, DOI: 10.1142/S0219622006002258.
9. Sun Y., Wong A.K., Kamel M.S.: "Classification of imbalanced data: A review", International Journal of Pattern Recognition and Artificial Intelligence, 2009, vol. 23(4), pp. 687–719, DOI: 10.1142/S0218001409007326.
10. Agnieszka Wosiak, Sylwia Karbowiak: “Preprocessing Compensation Techniques for Improved Classification” Proceedings of the Federated Conference on Computer Science and Information Systems pp. 203–211, Vol.11 IEEE 2017.
11. Retrieved from: https://bambielli.com/til/2017-10-29-gini-impurity/#
12. [12]. Retrieved from: https://dinsdalelab.sdsu.edu/metag.stats/code/randomforest.html.
13. Retrieved from: https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/
14. Retrieved from: https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/
15. Chawla N., Bowye K., Hall L., Kegelmeyer W.P.: "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, 2002, vol. 16, pp. 321â˘AS¸357, DOI: 10.1613/jair.953.
16. Retrieved from: https://machinelearningmastery.com/feature-selection-machine-learning-python/
17. Retrieved from:m-asim.com.
18. Xiaoying Guo, Yuhua Qian, Liang Li, Akira Asano. "Assessment model for perceived visual complexity of painting images", Knowledge-Based Systems, 2018.

AUTHORS PROFILE

Dr.K.Nandhini received her B.Sc., from Bharathiar University, Coimbatore in 1996 and received M.C.A from Bharathidasan University, Trichy in 2001. She obtained her M.Phil. in the area of Data Mining from Bharathidasan University; Trichy in 2004.She obtained her Ph.D degree in the area of data mining from Bharathiar University, Coimbatore in 2012. She is having 19 years of experience in academic and research. At present she is working as an Assistant Professor at PG & Research Department of Computer Science in Chikkanna Government Arts College, Tirupur. She has produced more than 6 M.Phil research scholars and published more than 25 research papers in various National and International journals(includes Scopus indexed) in the area of Data Mining. She acted as a resource person in various seminars and conferences. She delivered lot of lectures in various colleges of Tamilnadu. Her research interest lies in the area of Data Mining and Artificial Intelligence. She is a life member of the professional body Computer Society of India (CSI).

Mrs.V.Shobana is a Ph.D Research Scholar in PG and Research Department of Computer science, Chikkanna Government arts college, Tirupur. Also she is working as an Assistant Professor in Department of Computer Science, Dr.N.G.P. Arts and Science College, Coimbatore. She has published more than 8 papers in various reputed journals (includes Scopus indexed). Her areas of interest are data mining, big data analytics and R Programming.