ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Efficient Imbalanced Data Handling Techniques through Undersampling and Oversampling Approach

Mayuri S. Shelke, Dr. Prashant R. Deshmukh, Prof. Vijaya K. Shandilya

Student of Master of Engineering in (CSE), Sipna College of Engineering and Technology, Amravati, India

Professor, Sipna College of Engineering and Technology, Amravati, Maharshtra, India

Associate Professor, Sipna College of Engineering and Technology, Amravati, India

ABSTRACT: As datasets are of most important that gives you lots amount of varieties in single data file, which helps for number of activities like decision making, grouping similar type of records, generating predictions and recommendations, etc. Some of these datasets related to medical, e-commerce, social networking, etc. are of great importance. But many of these datasets are imbalanced in nature that is some records belonging to same category are in much large number and some are very rare. For extracting useful date from such large dataset different data mining or machine learning techniques are used. But due to the imbalance nature of dataset, the classifier unable to classify it properly or correctly. This problem refers to the uneven distribution of instances among the classes which poses difficulty in the classification of rare instances. To deal with this it is necessary to understand the problem of imbalanced learning and apply methods to deal with this. There are various Undersampling and oversampling techniques available which try to resolve imbalanced learning problem. To this end, the methods are proposed both for undersampling and oversampling. Majority class samples can be undersampled using a new approach, namely, MLP-Based Undersampling and Majority Weighted Minority Oversampling Technique (MWMOTE) can be used for generating the synthetic samples for minority class. In the develop system, the main objective is to handle the imbalance classification problem occurring in the medical diagnosis of rare diseases and combines the benefits of both undersampling and oversampling.

KEYWORDS: Imbalanced learning, Undersampling, Oversampling, MLP-Based Undersampling, Majority Weighted Minority Oversampling Technique (MWMOTE).

I. INTRODUCTION

In today's era of machine learning and data mining, many real world applications work on datasets mainly for performing analysis and generating recommendations and predictions. For performing these calculations the dataset should be properly balance [1]. But sometimes it is seen that these datasets are imbalance in nature. Leading to the problem of imbalanced data. The data which has an unequal distribution of samples among classes is known as imbalanced data. The class having more samples is generally a majority class and a class which contains very scarce samples is a minority class. Such type of data sets pose a great challenge to the classifier as it becomes problematic to classify the minority samples precisely because of their fewer amount [2]. Standard classification algorithms also fail to classify such form of imbalanced data accurately with least misclassification error.

Oversampling and undersampling in data analysis are techniques used to adjust the class distribution of a data set. Oversampling and undersampling are opposite and roughly equivalent techniques. They both involve using a bias to select more samples from one class from another [3]. There are major methods available to resolve the imbalanced learning problems which are nothing but a sampling, active learning, cost sensitive learning and kernel based methods. Sampling based methods provide the solution at data level by balancing the number of samples among classes.

DOI: 10.15680/IJIRCCE.2017. 0504349

Undersampling and oversampling are two main methods of sampling in which samples are either reduced from majority class or samples are added in the minority class. Both techniques have their own advantages as well as drawbacks [4]. Active learning approaches focus mainly on acquiring labels to the unlabeled data. Another method is cost based method which provides solution to an imbalanced dataset at the algorithmic level. It uses cost matrix which represents costs associated with each representation. Besides of these methods, kernel based methods also work well in handling imbalanced datasets [5].

Most of the machine learning algorithms perform better when data sets are almost balanced. If the datasets are imbalance in nature, classification of these imbalanced datasets is a very crucial task for the classifier as classifier may tend to favor the majority class samples, results in unequal distribution of data [6]. So, to deal with problems that are arises when given data sets are very much imbalanced in nature, this paper proposes the methods for both Undersampling and Oversampling. For Undersampling here MLP-based undersampling technique is used which will preserve the distribution of information while doing undersampling [7]. And for Oversampling Majority Weighted Minority Oversampling Technique (MWMOTE) is used. The objective of MWMOTE is to improve the sample selection process and to improve the synthetic sample generation process [8]. The paper propose the solution and find out the result generated obtained by the solution. Finally, it will also performs the analysis of result generated by the proposed solution.

II. LITERATURE SURVEY

To handle the imbalanced learning problem significant works have been done which can be categorized as a: sampling based methods, cost based methods, active learning based methods and kernel based methods [1]. Various sampling methods are done for undersampling and oversampling. Reduction of samples from majority class is done in undersampling. These methods again divided into random and informed undersampling. Random undersampling technique arbitrarily eliminates the samples from majority class, but it may lead to loss of important samples also. To overcome this issue, researchers proposed some informed undersampling techniques such as EasyEnsemble, Balancecascade [2], and KNN based methods, namely Near miss 1, Near miss 2, Near miss 3 and most distant method. One sided selection method also performs well to deal with imbalanced data. To further refine this method cluster-OSS has been proposed.

In oversampling techniques, artificial instances are added to the positive class (minority) to balance between classes. Oversampling can be random or synthetic sample generation. In random oversampling, samples are randomly replicated, but which can lead to over fitting. On the other hand, in synthetic oversampling method, it generates the synthetic samples to minority class. These generated samples add essential information to the minority class, resulting in improved performance of the classifier. In [2] proposed a powerful method, namely synthetic minority oversampling technique (SMOTE) which has been applied successfully in many applications.

III. METHODOLOGIES USED

A. MLP based Undersampling:

Undersampling methods work by reducing the majority class samples. This reduction can be done randomly in which case it is called random undersampling or it can be done by using some statistical knowledge in which case it is called informed undersampling. Some informed undersampling methods and iteration methods also apply data cleaning techniques to further refine the majority class samples. In this paper, MLP-based undersampling techniques are proposed to reduce the majority class samples. The complete implementation of MLPUS methodology is shown the figure 1 below [9].

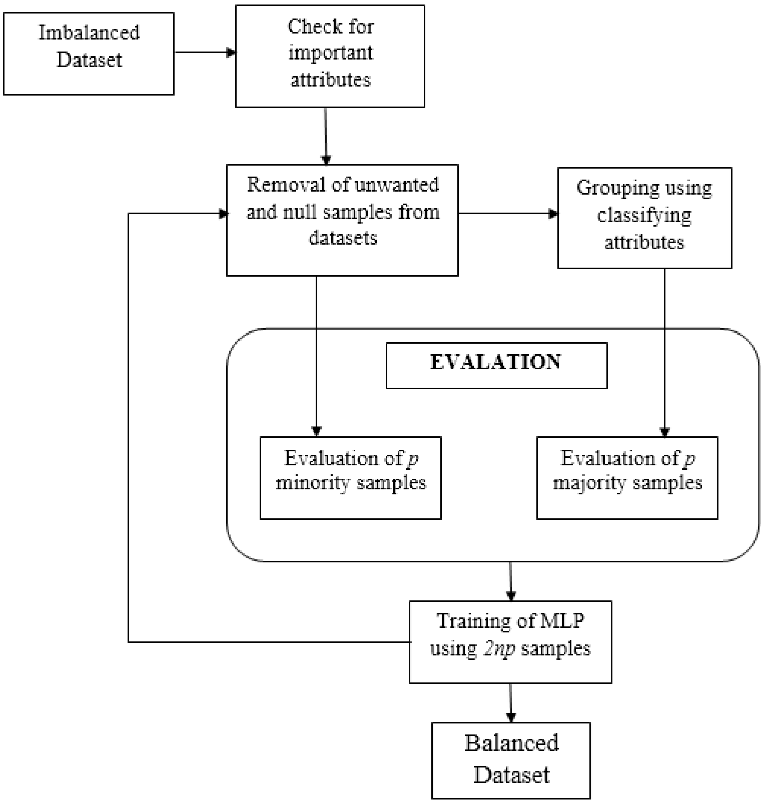
Imbalanced Dataset
Check for important attributes
Removal of unwanted and null samples from datasets
Grouping using classifying attributes
EVALATION
Evaluation of p minority samples
Evaluation of p majority samples
Training of MLP using 2np samples
Balanced Dataset

Figure 1: MLPUS Implementation

Mathematical model relevant to the system:

Notations:
S_{maj}: Set of majority class samples
S_{min}: Set of minority class samples
T_p: Total number of samples
Count: General count variable, initially set to zero.

Problem Description:
Let S be the system,
S= S_{maj}, S_{min}
Where, $S_{maj} > S_{min}$

Undersampling with MLPUS:
Step 1: Training the initial MLP

1) Make Group of both S_{maj} and S_{min} from total number of samples T_p
2) Set both P0 and R0 to be empty sets, add samples of two categories in P_0 and R_0
3) For each group of the minority class, if $P_0 > R_0$ set $S_{min} = R_0$
4) For each group of the majority class, if $R_0 > P_0$ set $S_{maj} = P_0$
5) $T_p = P_0 \text{ U } R_0$

Step 2: Find most important samples from S_{maj}
Step 3: Creates the groups of Samples S from S_{maj}
Step 4: Train a MLP using S

Step 5: Display Match that is sets S = S_{maj} = S_{min}

B. Oversampling using MWMOTE:

In oversampling method, new samples are added to the minority class in order to balance the data set. For oversampling mainly two methods are used called as random oversampling and synthetic oversampling. In random oversampling method, existing minority samples are replicated in order to increase the size of a minority class. But here in this technique there is chances of important samples becomes rare and less important attributes may be replicated. That's why, this paper uses the technique of MWMOTE based on synthetic oversampling. In this method artificial samples are generated for the minority class samples. These new samples add the essential information to the minority class and prevents its instances from the misclassification. The complete implementation of MWMOTE methodology is shown the figure 2 below [10].

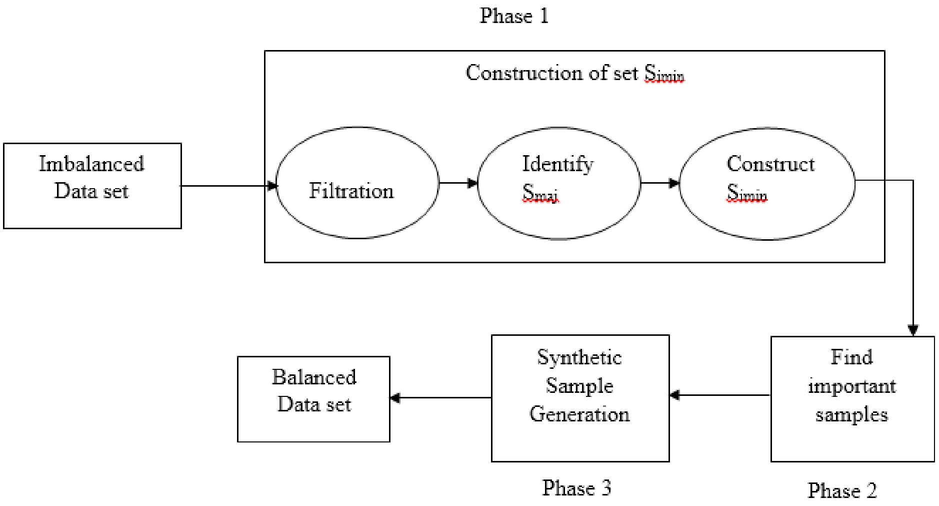

Figure 2: MWMOTE Implementation

MWMOTE implementation involves three key phases.

- In the first phase, MWMOTE identifies hard-to-learn and the most important minority class samples from the original minority set S_{min} and construct a set S_{imin} by the identified samples.
- In the second phase, selection of important samples from S_{imin} is performed, according to its importance in the data.
- In the third phase, MWMOTE generates the synthetic samples from Simin using important attributes obtained in the second phase and produces the output set, by adding the synthetic samples to Smin. This results in balanced data set.

IV. IMPLEMENTATION OF THE SYSTEM

A. Datasets Used

Any data sets for medical diagnosis can be taken as an input to the system. These data sets should have imbalanced nature. And all data sets should be in the binary form, means it should contain exactly two classes, one is majority and another is minority. The data sets can be of any size, and having different attributes and different imbalance ratios. In this paper, following five data sets are used by the system and that are taken from UCI repository.

- Pima Diabetes
- Breast Cancer

- Hepatitis
- Liver Disorder
- Heart Diseases

Stepwise implementation for data sets:

Step 1: Take dataset from UCI repository
Step 2: Read and understand relevant information
Step 3: Read and understand attribute information and select important attributes useful for our proposed work
Step 4: Convert the dataset file in CSV (Comma separated values) including only useful columns for our task. // (skip conversion if already exist in csv format).
Step 5: Create database containing same number of column and type as like attribute selected from dataset for calculations.
Step 6: Import the CSV file into MySQL database.

In sample selection of both minority and majority class, firstly noisy samples would be removed. Then samples which are important in the data set would be selected in sample generation process. Using these selected samples synthetic samples will be generated for selected minority samples or removed from majority class. The complete step are given below.

B. System Workflow:

There are mainly 5 use cases in the system. They are as follows:

- Input data set:
  Initially user will give imbalanced data set of rare diseases to the system to make it balance for medical diagnosis purpose.
- Apply MLPUS:
  It is used to do the undersampling of data set given by the user. It includes important attribute selection and grouping of data accordingly. Then it is provided as an input to train MLP. In MLP training samples are trained iteratively to avoid the misclassification.
- Apply MWMOTE:
  Majority weighted minority oversampling technique will take imbalanced data as an input and generate synthetic samples for minority class using groups of important attributes. It uses selection of important samples according to its importance in datasets, and generate the new samples accordingly.
- Classify:
  As MLPUS minimizes the samples of majority class and MWMOTE increases number of samples in the minority class, these both techniques balance the data. After balancing the datasets, if the is given to any classifiers it will give more proper and accurate results.
- Evaluation:
  After classification, the user can evaluate the results on MLPUS and MWMOTE separately as well as combining them together. This evaluation can be done on the basis of different parameters like precision, recall, G-Mean, overall accuracy and number of samples under consideration.

V. RESULT ANALYSIS

The goal of the proposed system is to handle the imbalanced learning problem arises in the medical diagnosis of rare diseases. It will take imbalanced data of rare diseases as an input and produces balanced data. This system will contain two major subsystems. One is used for reducing the samples from majority class and another is used to increase the number of samples in the minority class. Both these subsystems produce different results of balanced data sets. In this way data sets can be balanced and that can be used further for applying any data mining techniques on them.

Table 1: Performance measures of MLPUS

DataSet	Precision	Recall	G-Mean	Overall Accuracy
Heart_Diseace	0.219338	0.749778	0.40553	62.15
Breast Cancer	0.344778	0.749689	0.508406	55.24
Hepatitis	0.896774	0.749461	0.819815	84.63
Liver_Disorder	0.392405	0.749516	0.542323	94.27
Pima_Diabetes	0.348958	0.74972	0.511489	85.24

To evaluate the performance of proposed work various performance measures can be derived such as precision, recall, overall accuracy and G-mean. Table 1 show all these measures for MLPUS with respect to five datasets taken here as input to the system. And the table 2 shows the performance parameters of MWMOTE in tabular format as shown below.

Table 2: Performance measures of MWMOTE

DataSet	Precision	Recall	G-Mean	Overall Accuracy
Heart_Diseace	0.197404	0.949719	0.432988	81.17
Breast Cancer	0.3103	0.949606	0.542829	99.27
Hepatitis	0.807097	0.949317	0.875323	78.72
Liver_Disorder	0.353165	0.949388	0.579042	85.31
Pima_Diabetes	0.314063	0.949646	0.546121	33.28

Finally, we have evaluated our result obtained between our proposed techniques that is between undersampling and oversampling. The result obtained for all the parameters for both MLPUS and MWMOTE is shown below in graphical format in figure 3 below.

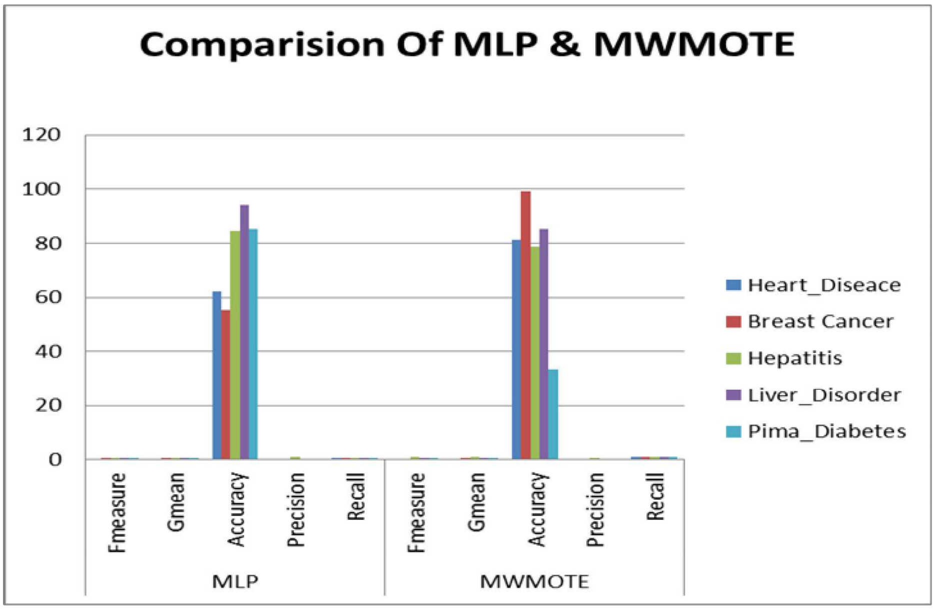

Figure 3: Graph shows comparison of MLPUS with MWMOTE

DOI: 10.15680/IJIRCCE.2017. 0504349

VI. CONCLUSION

Currently, imbalanced learning becomes a challenging and active research topic in the area of data mining and machine learning. So, to handle this imbalanced learning problem, here the system uses MLPUS as undersampling and MWMOTE as oversampling techniques. These techniques can be generalized to solve multiclass imbalanced learning problem. Furthermore, these methods can also be modified to facilitate for incremental learning applications and balancing the datasets. Here, the developed system is evaluated with number of different parameters and different techniques available for oversampling and undersampling and the result obtained is shown in both tabular and graphical format. It is seen that, the propose methods overcomes problems face by existing system. And it is desirable to develop a solution which will integrate benefits of both undersampling and oversampling and handles the imbalanced data in the medical diagnostic field efficiently.

REFERENCES

[1] Varsha Babar, Roshani Ade, "A Novel Approach for Handling Imbalanced Data in Medical Diagnosis using Undersampling Technique", *Communications on Applied Electronics (CAE), Foundation of Computer Science FCS, New York*, USA -2015.

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

[3] Yong Hu, Dongfa Guo, Zengwei Fan, Chen Dong, Qiuhong Huang, Guifang Liu, Boping Li, Qiwei Xie, "An Improved Algorithm for Imbalanced Data and Small Sample Size Classification", *Journal of Data Analysis and Information Processing*, 3, 27-33 Published Online August 2015. in SciRes. http://dx.doi.org/10.4236/jdaip.2015.33004

[4] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning", ICIC 2005, Part I, LNCS 3644, pp. 878 – 887,© *Springer-Verlag Berlin Heidelberg* 2 005.

[5] Bee Wah Yap, Khatijahhusna Abd Rani, Hezlin Aryani Abd Rahman, Zuraida Khairudin, Nik Nairan Abdullah, "An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets", *Proceedings of the First International Conference on Advanced Data and Information Engineering*, _ Springer Science + Business Media Singapore 2014.

[6] S. J. Yen and Y. S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Systems with Applications*, vol. 36, pp. 5718–5727, 2009.

[7] Alexander Yun-chung Liu, B.S, "The Effect of Oversampling and Undersampling on Classifying Imbalanced Text Datasets", The University of Texas at Austin, August 2004.

[8] H. He and E.A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge Data Eng., vol. 21, no. 9, pp. 1263-1284, Sept. 2009.

[9] Varsha Babar, Roshani Ade, "MLP-Based Undersampling Technique for Imbalanced Learning".

[10] Sukarna Barua, Md. Monirul Islam,Xin Yao, "MWMOTE-Majority Weighted Minority Oversampling Technique for imbalanced data set learning", IEEE Trans. Knowledge anddata engineering, vol. 26, no. 2, February 2014.