International Journal of Research in Advent Technology, Special Issue, March 2019
E-ISSN: 2321-9637
International Conference on Technological Emerging Challenges (ICTEC-2019)
Available online at www.ijrat.org

A Comparative Study Of Skewed Data Sources Using Fusion Sampling Using Diversified Distribution

[1]G. Shobana, [2]Dr Bhanu Prakash Battula
[1]*Rajiv Gandhi University of Knowledge Technology, College in Nuzvid, India.*
[2]*Tirumala Engineering College , Narasaraopet, India.*
Email: shobana.gorintla@gmail.com, prakashbattula33@gmail.com

Abstract: In data mining, imbalance learning is a challenging task due to the intrinsic properties of the imbalance datasets. An imbalance data consists of unequal ratio instances in the classes. To address the limitations of imbalance data, we compared different novel algorithm taking into account both under sampling and over sampling. In fact, our algorithm is capable of restructuring the original dataset at a very high conceptual level to alleviate the problems in the class imbalance. We conduct the empirical benchmark experimental setup using 9 datasets of varying class imbalance level. The experimental results suggest that the proposed approach performs effectively than the existing approaches.
Keywords: Data Mining, Knowledge Discovery, Classification, oversampling, under sampling, Fusion Sampling, Diversified Distribution.

1. INTRODUCTION

Decision trees are the mathematical based algorithmic model which uses logic as the core unit for decision making. Decision tree consists of the branches and leaves. Each branch is a path of splitting the records into a narrow space and each leaf is the result of the classification of records in a specific class. There are numerous models of decision trees, which access the data and classify them in the predefined classes.

Rukshan Batuwita et al., [1] have reviewed to conclude that SVMs could produce suboptimal models which are biased towards the majority class and have low performance on the minority class. Rushi Longadge et al., [2] have gathered the evidence to show that the most of algorithm are more focusing on classification of major sample while ignoring or misclassifying minority sample when imbalance dataset are applied. Kun Jiang et al., [3] have propose a novel genetic algorithm-based SMOTE (GASMOTE) algorithm which uses different sampling rates for different minority class instances and finds the combination of optimal sampling rates.

Shaza M. Abd Elrahman et al., [4] have reviewed a general survey for class imbalance problem solutions and the most significant researcher's investigations. Bartosz Krawczyk [5] has provided a discussion and suggestions concerning lines of future research for classification, regression, clustering, data streams and big data analytics.

The review of the recent works suggests that the efficiency of the decision tree reduces drastically when applied for class imbalance data sources. The reason for the reduction in performance is due to the inefficient model built with the rare instances class.

The arrangement of paper is follows as. We exhibit in section 2 the recent approaches in learning with decision tree. It will straightforwardly persuade the principle commitment of this work introduced in section 3. We compared another structure for improved learning. Assessment criteria's designed for decision tree learning is exhibited in section 4. Test results are accounted in section 5. In conclusion, we finish up with section 6 where we talk about real open issues and upcoming work.

2. RELATED WORK:

Chongsheng Zhang et al [6] have reviewed a first empirical study on the performance of the two opposing pipelines for binary imbalance learning, i.e., first feature selection then resampling, or first resampling then feature selection. Shuo Wang et al [7] have performed a systematic study of handling concept drift in class-imbalanced data streams; including current research focuses and open challenges. Shuo Wang et al [8] have studied the issue of if and how class imbalance learning methods can benefit software defect prediction with the aim of finding better solutions. They investigated different types of class imbalance learning methods, including resampling techniques, threshold moving, and ensemble algorithms.

Lov Kumar et al [9] have conducted a study on the application of static source code metrics and machine learning techniques to predict aging related bugs in class imbalance software engineering datasets. Shuo Wang et al [10] have studied the combined challenges posed by multiclass imbalance and online learning, and aims at a more effective and adaptive solution. They introduced two resampling-based ensemble methods, called MOOB and MUOB, which can

process multi-class data directly and strictly online with an adaptive sampling rate. M. Mostafizur Rahman et al [11] have examined the performance of over-sampling using SMOTE and an improved under-sampling technique to balance cardiovascular data.

Amritanshu Agrawal et al [12] have applied a multi-performance criteria's AUC and recall while fixing the weaker regions of the training data using SMOTUNED, which is an auto-tuning version of SMOTE. Jianhong Yan et al [13] have proposed a novel RE-sample and Cost-Sensitive Stacked Generalization (RECSG) method based on 2-layer learning models. The first step is Level 0 model generalization including data pre-processing and base model training. The second step is Level 1 model generalization involving cost-sensitive classifier and logistic regression algorithm. Bo SUN et al [14] have introduce an under-sampling bagging framework and proposed an evolutionary under-sampling (EUS) based bagging ensemble method EUS-Bag by designing a new fitness function considering three factors to make EUS better suited to the framework.

Sudarsun Santhiappan et al [15] have proposed a novel unsupervised topic modelling based weighting framework to estimate the latent data distribution using a topics oriented directed under-sampling algorithm that follows the estimated data distribution to draw samples from the dataset. Siqi Ren et al [16] have proposed an ensemble classifier called Gradual Resampling Ensemble (GRE), which handles data streams using a selectively re-sampling method, where drifting data can be avoidable, is applied to select a part of previous minority examples for amplifying the current minority set which exhibit concept drifts and class imbalance. Khaldy MA et al [17] have explored and analysed different feature selection methods to select a subset of the original data and then resample for a clinical dataset that suffers from high dimensional and imbalance data.

M. Muksitul Haque et al [18] have investigated a number of imbalanced class algorithms including the TAN+ AdaBoost algorithm for solving the imbalanced class distribution present in epigenetic datasets which inherently come with few differentially DNA methylated regions (DMR) and with a higher number of non-DMR sites. For this class imbalance problem, a number of algorithms are compared. Neelam Rout et al [19] have discussed the meaning of the imbalanced data, examples of the imbalanced data, different challenges of handling the imbalanced data, imbalance class problems and performance analysis metrics for the imbalanced data are elaborated in different scenario.

Georgios Douzas et al [20] have proposed a conditional version of Generative Adversarial Networks (cGAN) to approximate the true data distribution and generate data for the minority class of various imbalanced datasets to validate against multiple standard oversampling algorithms. Samir Al-Stouhi et al [21] have developed a method that is optimized to simultaneously augment the training data and induce balance into skewed datasets. They proposed a novel boosting-based instance transfer classifier with a label-dependent update mechanism that simultaneously compensates for class imbalance and incorporates samples from an auxiliary domain to improve classification. Brendan Juba et al [22] have consider the measures of classifier performance in terms of precision and recall, a measure that is widely suggested as more appropriate to the classification of imbalanced data. They observed that whenever the precision is moderately large, the worse of the precision and recall is within a small constant factor of the accuracy weighted by the class imbalance ad the solution is that the only cure for class-imbalance is a larger number of examples.

3. THE PROPOSED FRAMEWORKS:

3.1 Framework of Over Sampled Diversified Distribution (OSDD) Algorithm

This section presents the proposed algorithm Over Sampling using Diversified Distribution (OSDD), whose main characteristics are depicted in the following sections.

Some researchers have already shown that the global imbalanced ratio between classes is not a problem itself and it may not be the main source of difficulties [22]. The degradation of classification performance is linked to other factors related to data distribution, such as decomposition of the sub classes into many rare sub-concepts. This problem of distribution of data can be termed as diversified distribution. In this paper, we propose to use the concept of diversified data distribution especially in the minority class to perform oversampling.

In a binary class imbalance data, the class with high percentage of instances is called as the majority class and the class with less percentage of instances is known as minority class. The oversampling of the instances in the minority class will reduce the problem of class imbalance. The improper oversampling of data in the minority class can lead to disaster of the data source with an irreparable state. A proper statistical or technical analysis is to be conducted before initiating the oversampling process. In our work, the concept of diversified data distribution is used to study and identify different regions in the minority subset.

The identified regions are categories into safe, borderline and unsafe regions. An action should be initiated for borderline and unsafe regions whereas no action is required for the safe regions. The unsafe regions are the regions formed with the noisy or outlier instances. The removing of the unsafe regions will help to improve the quality of minority subset. The borderline regions are formed with a mix of instances which belong to majority and minority subsets. The improvement of these regions is required for refining of the minority subset.

The oversampling of the reaming stronger instances in the minority subset is initiated. The percentage of oversample can range from 10-100% depending on the intrinsic properties of the dataset. In the final stage, the improve minority subset and majority subset are combined and applied to a base algorithm for results simulation. Here, we have considered best first decision tree [23] as the base algorithm for our frame work. A best-first decision tree classifier uses binary split for both nominal and numeric attributes. For missing values, the method of 'fractional' instances is used.

3. 2 Framework of Under Sampled Diversified Distribution (USDD) Algorithm

Based on current literature, characteristics of class imbalance datasets usually include overlapping classes, diversified distribution, sparsely populated, and outliers or noisy data [1-22]. Based on these characteristics, this paper proposed a novel under sampling method designed to addresses such problems. The theory of the proposed methodology will be first demonstrated using hypothetical datasets displaying these characteristics and evaluated using real-world datasets from the UCI data repository [23].

A. Identification of Majority Data Space

The proposed Under Sampled Diversified Distribution (USDD) algorithm is implemented via a two step strategy. The motivation is to accurately identify the majority data space and reducing the unwanted instances with quick and able techniques. Therefore, the first step in this procedure is to make use of clustering methods on majority data and identify weak and mini clusters. This set of weak and mini clusters represents the unnecessary data space of the majority space.

In order to cater for a variety of datasets, whereby the majority data space can take on various shapes and characteristics, multiple clustering methods are implemented. The simple k-means clustering method is chosen in this paper while maintaining scalability toward large datasets in order to ensure suitability for a variety of datasets. However, the choice of clustering method is not limited; any number of alternative clustering methods may be chosen based on characteristics of datasets in the field of application.

B. Cluster reduction and Outlier Removal:

It should be noted that the USDD method has a unique property of determining the majority data space using a threefold mechanism. The first mechanism is the choice of the clustering method which would cater for different data shapes. Second, the choice of the number of clusters (*k*) breaks up the minority data space into smaller pieces, catering for diversified distribution in the data. Finally, the value of nearest neighbour has an effect of merging clusters or outlier removal. Using a nearest neighbour value smaller than the number of data points within a cluster eliminates circumferential data points from the majority space. However, a nearest neighbour value larger than the number of data points within a cluster would allow inter cluster deletion of data points, thus decreasing the clusters. These three properties allow USDD to fluidly alter the region of data reduction to best match the original data space.

C. Merging the minority and the improved majority space

Once the data space has been accurately improved, the final step is to combine the minority ad improved majority data space. The proposed USDD method uses random forest [24] as the foundation algorithm for results interpretation. However, the novelty of this method is choice of data points which undergoes interpolation for elimination.

3. 3 Framework of Fusion Sampling using Diversified Distribution (FSDD) Algorithm

This section presents the detail architecture of the proposed Fusion Sampling using Diversified Distribution (FSDD) approach which consists of four major modules. The detailed working principles of the FSDD approach are explained below in the sub-sections.

In the initial stage of our frame work the dataset is divided into minority subset P € pi (i = 1,2,..., pnum) and majority subset N € ni (i = 1,2,..., nnum) respectively. The minority subset is the class of instances which are very less when compared to the other class in the dataset. The majority subset is the class of instances, which are more in percentage than the other class.

As the traditional algorithms efficiency drops down on imbalance data, to improve the efficiency, the dataset's majority subclass is to be under sampled or minority subclass is to be oversampled. In our proposed approach we initiated the both under sampling and oversampling strategy for the majority and minority sub classes respectively. One

of the limitations of the existing oversampling algorithms is of not considering for removal of noisy and outlier instances before oversampling. Therefore, in the proposed approach before oversampling phase is started mostly misclassified instances are removed from the dataset in the form of under sampling. The technique proposed for identifying the mostly misclassified instances is by considering the nearest neighbor instances. If all the nearest neighbor instances of a particular instance are of opposite class then it implies that particular instance comes under the category of a noisy or outlier instance and can be eliminated.

The instances in the majority subset are reduced by following the below mentioned techniques; one of the technique is to eliminate the noise instances, the other technique is to find the outliers and the final technique is to find the range of weak instances for removal. The noisy and outlier instances can be easily identified by analyzing the intrinsic properties of the instances. The range of weak instances can be identified by first identifying the weak features in the majority subset. The correlation based feature selection [23] technique selects the important features by following the inter correlation between feature - feature and the inter correlation between feature and class. The features which have very less correlation are identified for elimination. The range of instances which belong to these weak features are identified for elimination from the majority subset. The number of features and instances eliminate by the correlation based feature selection technique will vary from dataset to dataset depending upon the unique properties of the dataset. The eliminated instances can boost the performance of the proposed approach in two ways:

First it will reduce the noisy and outlier instances not only from majority but also minority subset and hence improves the quality of the dataset. Second it reduces some of the outlier and noisy instances from majority subset and so reduces the imbalance nature of the dataset.

In the next phase minority subset is oversampled. The some of the synthetic instances generated are the replica of the existing instances, hybrid instances and pure artificial instances.

In the final stage the fine tuned dataset is applied to base algorithm here random forest [24] is considered and evaluations metric are generated.

4. INVESTIGATIONAL DESIGN AND ASSESSMENT CRITERIA'S

The details of the datasets are given in table 1. For each data set, S.no., Dataset, name of the dataset, Instances, number of instances, Attributes, number of attributes, IR, imbalance ratio are described in the table for all the datasets. The most popular machine learning publicly available datasets are available at Irvine [25].

Table 1 The UCI datasets and their properties

S.no. Dataset	Inst	Attributes	IR
1. Breast	286	9	2.37
2. Breast_w	699	9	1.90
3. Colic	368	22	1.71
4. Credit-g	1,000	20	2.33
5. Diabetes	768	8	1.87
6. Hepatitis	155	20	3.85
7. Ionosphere	351	35	1.79
8. Labor	57	17	1.85
9. Sonar	208	13	1.15

The evaluation metrics used in the paper are detailed below, the percentage of instances correctly classified by a classified is known as Accuracy. AUC can be computed simple as the micro average of TP rate and TN rate when only single run is available from the clustering algorithm. The AUC is defined as the mean of true positive rate and true negative rate. The formula for AUC is given below,

$$AUC = \frac{1 + TP_{RATE} - FP_{RATE}}{2} \quad \text{..... (1)}$$

Or

$$AUC = \frac{TP_{RATE} + TN_{RATE}}{2} \quad \text{..... (2)}$$

The Precision measure is computed by,

$$\Pr ecision = \frac{TP}{(TP)+(FP)} \quad \ldots\ldots\ldots\ldots (3)$$

The Recall measure is computed by,

$$\mathrm{Re}call = \frac{TP}{(TP)+(FN)} \quad \ldots\ldots\ldots\ldots\ldots\ldots$$

(4)

The F-measure value is computed by,

$$F-measure = \frac{2\times \Pr ecision\times \mathrm{Re}call}{\Pr ecision+\mathrm{Re}call}$$

…. (5)

5. RESULTS

In this section, the results of the various approaches are compared and discussed. In order to test the strength of our method compared to existing methods, we included Under Sampling and over sampling in our experiments and the implementation algorithm in the java programming language using the open source tool weka [26]. We evaluated each of the classifiers on the eleven datasets from a number of different resources of UCI data repositories (Table 1). The results are summarized as follows.

Table 2 Summary of tenfold cross validation performance for AUC on all the datasets

Datasets	OSDD	USDD	FSDD
Breast	0.774±0.083	0.652±0.118	**0.976±0.021**
Breast_w	0.971±0.022	0.989±0.011	**0.998±0.004**
Colic	0.928±0.045	0.900±0.056	**0.985±0.013**
Credit-g	0.800±0.040	0.733±0.054	**0.996±0.006**
Diabetes	0.834±0.045	0.802±0.052	**0.991±0.008**
Hepatitis	0.897±0.089	0.851±0.126	**0.986±0.029**
Ionosphere	0.939±0.045	0.972±0.034	**1.000±0.000**
Labor	0.930±0.094	0.958±0.102	**0.993±0.024**
Sonar	0.815±0.106	0.893±0.068	**0.995±0.010**

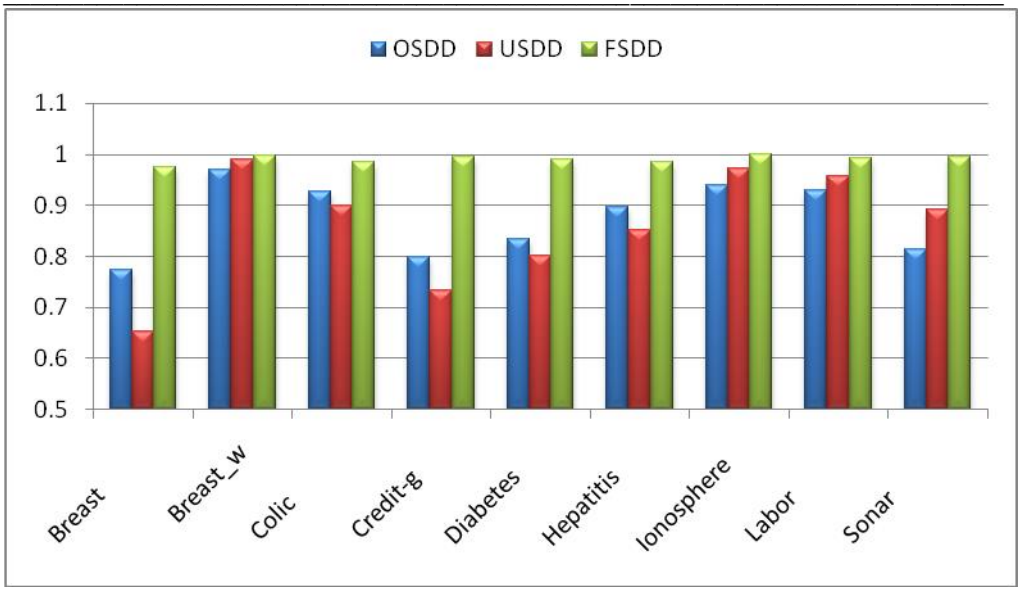

Fig. 1 Trends in AUC for OSDD, USDD versus FSDD on UCI data sets

Table 3 Summary of tenfold cross validation performance for Precision on all the datasets

Datasets	OSDD	USDD	FSDD
Breast	0.766±0.069	0.732±0.058	**0.951±0.032**
Breast_w	0.979±0.024	0.971±0.023	**0.995±0.008**
Colic	0.935±0.052	0.861±0.068	**0.958±0.026**
Credit-g	0.838±0.048	0.759±0.026	**0.985±0.014**
Diabetes	0.843±0.050	0.774±0.043	**0.977±0.018**
Hepatitis	0.729±0.178	0.639±0.358	**0.960±0.081**
Ionosphere	0.939±0.052	0.925±0.071	**0.992±0.017**
Labor	0.932±0.134	0.872±0.237	**0.974±0.081**
Sonar	0.822±0.102	0.794±0.094	**0.982±0.036**

Table 4 Summary of tenfold cross validation performance for Recall on all the datasets

Datasets	OSDD	USDD	FSDD
Breast	0.805±0.094	0.817±0.087	**0.956±0.040**
Breast_w	0.947±0.035	0.971±0.023	**0.995±0.009**
Colic	0.887±0.067	0.905±0.063	**0.970±0.028**
Credit-g	0.740±0.054	0.866±0.044	**0.982±0.014**
Diabetes	0.769±0.066	0.840±0.062	**0.979±0.019**
Hepatitis	0.804±0.189	0.435±0.263	**0.928±0.098**
Ionosphere	0.920±0.067	0.885±0.096	**0.991±0.020**
Labor	0.913±0.170	0.795±0.276	**0.948±0.109**
Sonar	0.848±0.114	0.822±0.118	**0.948±0.052**

Table 5 Summary of tenfold cross validation performance for F-measure on all the datasets

Datasets	RUS	USDD	FSDD
Breast	0.781±0.060	0.770±0.055	**0.953±0.026**
Breast_w	0.962±0.021	0.970±0.015	**0.995±0.006**
Colic	0.908±0.045	0.880±0.051	**0.964±0.019**
Credit-g	0.784±0.037	0.808±0.029	**0.983±0.010**
Diabetes	0.802±0.045	0.804±0.042	**0.978±0.013**
Hepatitis	0.744±0.143	0.487±0.257	**0.939±0.070**
Ionosphere	0.928±0.048	0.902±0.068	**0.991±0.013**
Labor	0.905±0.122	0.806±0.227	**0.955±0.078**
Sonar	0.828±0.084	0.802±0.082	**0.964±0.034**

Table 2 shows the detailed experimental results of the mean AUC of compared approaches. From Table 2 we can see AUC performance of FSDD model with a substantial improvement over other approaches suggests that the FSDD model is potentially a good technique for decision trees. The FSDD method can also gain significantly improvement over comparable to two state-of-the-art technique for decision trees.

Table 3 shows the detailed experimental results of the precision of FSDD approach with comparison. From Table 3 we can see FSDD model have performed well in terms of precision and have achieved substantial improvement over compared approaches. Table 4 shows the detailed experimental results of the recall of FSDD approach. From Table 4 we can see error reduction of FSDD model with a substantial decrease over other model is potentially a good technique for decision trees. The FSDD method has reduced error over other approaches. Table 5 shows the detailed experimental results of the mean F-measure of FSDD approach. From Table 5 we can see F-measure performance of FSDD model with a substantial improvement over compared approaches.

6. CONCLUSION

In this paper, we compared different approaches taking into account both under sampling and over sampling. In fact, our algorithm is capable of restructuring the original dataset at a very high conceptual level to alleviate the problems in the class imbalance. We conduct the empirical benchmark experimental setup using 9 datasets of varying class imbalance level. The experimental results suggest that the proposed approach performs effectively than the compared approaches.

In future work, we want to apply the proposed framework for multi class learning data sources.

REFERENCE:

[1] Rukshan Batuwita and Vasile Palade,"CLASS IMBALANCE LEARNING METHODS FOR SUPPORT VECTOR MACHINES", Imbalanced Learning: Foundations, Algorithms, and Applications,. By Haibo He and Yunqian Ma, Copyright c 2012 John Wiley & Sons, Inc.

[2] Rushi Longadge, Snehlata S. Dongre, Latesh Malik," Class Imbalance Problem in Data Mining: Review", International Journal of Computer Science and Network (IJCSN) Volume 2, Issue 1, February 2013 www.ijcsn.org ISSN 2277-5420.

[3] Kun Jiang, Jing Lu, Kuiliang Xia," A Novel Algorithm for Imbalance Data Classification Based on Genetic Algorithm Improved SMOTE", Arab J Sci Eng, DOI 10.1007/s13369-016-2179-2.

[4] Shaza M. Abd Elrahman and Ajith Abraham,"A Review of Class Imbalance Problem"Journal of Network and Innovative Computing ISSN 2160-2174, Volume 1 (2013) pp. 332-340 © MIR Labs, www.mirlabs.net/jnic/index.html

[5] Bartosz Krawczyk," Learning from imbalanced data: open challenges and future directions", Prog Artif Intell, DOI 10.1007/s13748-016-0094-0

[6] Chongsheng Zhang, Jingjun Bi, Paolo Soda," Feature Selection and Resampling in Class Imbalance Learning: Which Comes First? An Empirical Study in the Biological Domain", 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp.o: 933-938.

[7] Shuo Wang , Leandro L. Minku and Xin Yao," A Systematic Study of Online Class Imbalance Learning With Concept Drift", IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

[8] Shuo Wang and Xin Yao," Using Class Imbalance Learning for Software Defect Prediction, IEEE TRANSACTIONS ON RELIABILITY, VOL. 62, NO. 2, JUNE 2013.

[9] Lov Kumar, Ashish Sureka," Feature Selection Techniques to Counter Class Imbalance Problem for Aging Related Bug Prediction", ISEC '18, February 9–11, 2018, Hyderabad, India.

[10] Shuo Wang, Leandro L. Minku Xin Yao," Dealing with Multiple Classes in Online Class Imbalance Learning", Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16).

[11] M. Mostafizur Rahman and D. N. Davis," Addressing the Class Imbalance Problem in Medical Datasets", *International Journal of Machine Learning and Computing, Vol. 3, No. 2, April 2013.*

[12] Amritanshu Agrawal, Tim Menzies," Is "Better Data" Better Than "Better Data Miners"? On the Benefits of Tuning SMOTE for Defect Prediction", ICSE '18, May 27-June 3, 2018, Gothenburg, Sweden.

[13] Jianhong Yan and Suqing Han," Classifying Imbalanced Data Sets by a Novel RE-Sample and Cost-Sensitive Stacked Generalization Method", Mathematical Problems in Engineering Volume 2018, Article ID 5036710, 13 pages, https://doi.org/10.1155/2018/5036710.

[14] Bo SUN, Haiyan CHEN, Jiandong WANG and Hua XIE," Evolutionary under-sampling based bagging ensemble method for imbalanced data classification", Front. Comput. Sci. DOI 10.1007/s11704-016-5306-z

[15] Sudarsun Santhiappan, Jeshuren Chelladurai, and Balaraman Ravindran," A novel topic modeling based weighting framework for class imbalance learning", CoDS-COMAD '18, January 11–13, 2018, Goa, India

[16] Siqi Ren, Bo Liao, Wen Zhu, Zeng Li, Wei Liu, Keqin Li," The Gradual Resampling Ensemble for mining imbalanced data streams with concept drift", https://doi.org/10.1016/j.neucom.2018.01.063

[17] Khaldy MA, Kambhampati C (2018) Resampling Imbalanced Class and the Effectiveness of Feature Selection Methods for Heart Failure Dataset. Int

Rob Auto J 4(1): 00090. DOI: 10.15406/iratj.2018.04.00090

[18] M. MUKSITUL HAQUE, MICHAEL K. SKINNER, and LAWRENCE B. HOLDER," Imbalanced Class Learning in Epigenetics", JOURNAL OF COMPUTATIONAL BIOLOGY Volume 21, Number 7, 2014, Mary Ann Liebert, Inc. Pp. 492–507, DOI: 10.1089/cmb.2014.0008

[19] Neelam Rout, Debahuti Mishra and Manas Kumar Mallick, "Handling Imbalanced Data: A Survey", M.S. Reddy et al. (eds.), International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications, Advances in Intelligent Systems and Computing 628, https://doi.org/10.1007/978-981-10-5272-9_39

[20] Georgios Douzas , Fernando Bacao , Effective data generation for imbalanced learning using Conditional Generative Adversarial Networks, *Expert Systems With Applications* (2017), doi: 10.1016/j.eswa.2017.09.030

[21] Samir Al-Stouhi, Chandan K. Reddy," Transfer learning for class imbalance problems with inadequate data", Knowl Inf Syst, DOI 10.1007/s10115-015-0870-3.

[22] Brendan Juba, Hai S. Le," Precision-Recall versus Accuracy and the Role of Large Data Sets", Association for the Advancement of Artificial Intelligence (www.aaai.org)., 2018.

[23] Hall MA (1998) Correlation-based feature subset selection for machine learning. PhD Thesis.

[24] Leo Breiman (2001). Random Forests. Machine Learning. 45(1):5-32.

[25] Blake C, Merz CJ (2000) UCI repository of machine learning databases. Machine-readable data repository. Department of Information and Computer Science, University of California at Irvine, Irvine. http://www.ics.uci.edu/mlearn/MLRepository.html

[26] Witten, I.H. and Frank, E.(2005) Data Mining:Practical machine learning tools and techniques.2nd edition