International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 3, Issue 10, October 2014 **ISSN 2319 - 4847**

A survey on Oversampling Techniques for Imbalanced Learning

Miss Reshma K. Dhurjad[1], Prof. Mr. S. S. Banait [2]

[1] K. K. Wagh Institute of Engineering Education & Research, Department of Computer Engineering, Savitribai Phule Pune University

[2] K. K. Wagh Institute of Engineering Education & Research, Department of Computer Engineering, Savitribai Phule Pune University

ABSTRACT

In machine learning and data mining data imbalance is a key source of performance degradation. Key reason behind this degradation is that all available algorithms assume a balanced class distribution for learning. In many real-world applications, the data available for learning are highly imbalanced. Imbalanced data means where one class severely out-represent another class. In these scenarios, the learning algorithms tend to bias toward the less important negative class or majority class with larger instances. Although, there is no single best technique to deal with imbalance problems, sampling techniques have been shown to be very successful in recent years. To address imbalanced learning issue oversampling of minority class is done. There are various Oversampling techniques which can be used to reestablish the class balance. Oversampling method is a data level method. The main advantage of data level methods is that they are self-sufficient. The methods at data level modify the distribution of the imbalanced datasets, and then these modified i.e. balanced datasets are provided to the algorithm to improve the Imbalanced learning.

Keywords:- Classification, Imbalanced data, learning, oversampling

1. INTRODUCTION

Data Mining is the analysis of observational data sets to find unknown relationships and to summarize the data in novel ways that are both understandable and useful to the people [5]. There are many different data mining functionalities. Data characterization is the summarization of the general characteristics or features of a target class of data [6]. Machine Learning community, and in Data Mining works, Classification has its own importance. Classification is an important part and the research application field in the data mining [1]. With increasing volumes of operational data, many organizations have started to apply data-mining techniques to mine their data for new, valuable information that can be used to support their decision making [2]. Classification is the procedure of finding a set of models that distinguish and describe data classes, for the intent of being able to use the model to anticipate the class of objects whose class label is not known. Learning how to classify objects to one of a pre-specified set of categories or classes is a characteristic of intelligence that has been of keen interest to researchers in psychology and computer science. Identifying the common core characteristics of a set of objects that are representative of their class is of enormous use in focusing the attention of a person or computer program. Classification task can be done with two ways one is Supervised Learning and other is Unsupervised Learning. In supervised learning there is a specified set of classes, and example objects are labeled with the appropriate class. The goal is to generalize from the training objects that will enable novel objects to be identified as belonging to one of the classes. The second one is unsupervised learning. Often the goal in unsupervised learning is to decide which objects should be grouped together, in other words, the learner forms the classes itself. The success of classification learning is largely dependent on the quality of the data provided for training. A learner has only the input to learn from. If the data is inadequate then the concept descriptions will reflect this and misclassification will result when they are applied to new data. Data inadequacy is the measure cause of imbalance data. To learn from imbalance data is too difficult. Imbalanced learning problems contain unequal distribution of data instances among different classes, where most of the samples belong to some classes and rest to the other classes. If such instances come only from two classes, the class having most of the samples is called the majority class and other the minority class. Learning from the imbalance data is of uttermost important to the research community as it is present in many vital real-world classification problems, such as intrusion detection in network forensics, fraud detection in financial data analysis, cancer detection in biomedical diagnosis, object detection in computer vision, diagnosis and prognosis of machine failures, and so on.

2. PROBLEM OF IMBALANCED DATASETS

Imbalanced datasets means a dataset whose classification categories are not equally represented. The level of imbalance can be as large as 1:99[10]. It is notable that class imbalance is emerging as an crucial issue in designing classifiers [11], [12], [13]. Furthermore, the class with the few number of instances is usually the class of interest for the purpose of learning task [14]. This problem is of great interest because it turns up in many real-world classification problems, such as remote-sensing, pollution detection, risk management, fraud detection [18], and especially medical diagnosis [19]–[22]. For example, in case of earthquakes the known instances are rare but certainly of greater interest for earthquake prediction than the ample normal instances. In such situations positive class instances are sparsely distributed and negative class instances are densely distributed. In these scenarios, the learning algorithms tend to bias toward the less important negative class with larger instances.

3. TECHNIQUES TO DEAL WITH IMBALANCED DATASETS LEARNING:

Imbalanced datasets learning issue can be solve with two ways first is data level [3],[4],[7],[8],[9],[15] and the second one algorithmic levels [16],[17]. The techniques at data level alter the distribution of the instances in imbalanced data sets, and then it is given to the learner. The techniques at the algorithm level alter the present data mining algorithms or put up new algorithms to solve the imbalance problem. They enforce emphasis on the minority class by manipulating and incorporating learning parameters such as data-space weighting, class-dependent cost matrix, and receiver operating characteristics (ROC) threshold into conventional learning paradigms. The main advantage of data level methods is that they are self-sufficient. In this paper, we are laying more stress to study a data level oversampling methods for solving the class imbalance problem. To address imbalanced learning issue oversampling of minority class is done. To solve imbalanced learning issue, various oversampling methods were proposed like SMOTE [3], Borderline-SMOTE [8], ADASYN [9] SPO [24], INOS [25], DataBoost [23], so that a class balance is re-establish. Then the classifier is learn from the balanced dataset. This will definitely improve the efficiency of classification learning.

4. DATA LEVEL OVERSAMPLING TECHNIQUES FOR DATA BALANCING

In this section, an overview of oversampling techniques is provided. The objective of this survey is to clearly understand the Oversampling techniques to balance the distribution of data instances in the datasets. In year 2002 N.V. Chawla et al. proposed "SMOTE: Synthetic Minority Over-Sampling Technique". This work shows that a combination of method under sampling the majority class and oversampling the minority class can accomplish better classifier performance than only under sampling the majority class. Their method of oversampling the minority class includes creating synthetic minority class instances. SMOTE provides a new approach to oversampling. SMOTE and under-sampling in combination achieves better performance than plain under-sampling. The machine learning community has deal with the problem of class imbalance in two ways. First one is to assign unique costs to training instances. The second one is to resample the dataset, either by oversampling the minority class or under sampling the majority class. Authors approach combines under sampling of the majority class with a particular form of oversampling the minority class. SMOTE forces focused learning and introduce a bias towards the minority class. SMOTE classifier achieves better performance than Under-sampling classifier. SMOTE provides more related minority class instances to learn from, thus permit a one to carve broader decision areas, resulting in more coverage of the minority class. The SMOTE algorithm also has its drawbacks, including over generalization and variance. In the SMOTE algorithm, the issue of over generalization is mainly focused to the way in which it creates synthetic instances. Specifically, SMOTE gives the same number of synthetic data instances for each original minority instance and does so without consideration to neighboring instances, which enhances the occurrence of overlapping between classes. Various adaptive sampling techniques have been proposed to get over this limitation; some major methods includes the Borderline-SMOTE and Adaptive Synthetic Sampling (ADA- SYN) algorithms. Based on SMOTE method H. Han, W.Y. Wang et al.[8] were proposed two novel minority oversampling techniques, borderline-SMOTE1 and borderline-SMOTE2. In this only the minority instances near the borderline are oversampled. For the minority class, experiments show that borderline-SMOTE approach achieve better performance than SMOTE and random over-sampling methods. In order to achieve better prediction performance, most of the classification techniques attempt to learn the borderline of class as precisely as possible in the training process. The instances on the borderline and the ones nearby are more likely to be not categorized properly than the ones farthest from the borderline, and therefore more vital for classification. Based on the above discussion, those instances far from the borderline may contribute little to classification. Authors thus present two new minority oversampling techniques, borderline-SMOTE1 and borderline-SMOTE2. These techniques only oversampled borderline instances of the minority class. Their methods are different from the existing oversampling methods in which all the minority instances are oversampled. These methods are based on Synthetic Minority Oversampling method. SMOTE generates synthetic minority instances to oversample the minority class. For every minority instance, its *k* nearest neighbors of the each class are work out, then some instances are randomly chosen from

them according to the oversampling rate. After that, new synthetic instances are produce along the line between the minority instances and it's selected nearest neighbors. Not like the existing oversampling methods, these methods only oversample or strengthen the borderline minority instances. First, they find out the border line minority instances; then, synthetic instances are produce from them and added to the training set. Suppose that the training set is *T*, the minority class is *P* and the majority class is *N*, and

P= {p1, p2 ,..., p,pnum}, N= {n1,n2 ,...,n,nnum}

where pnum and nnum are the minority and majority instances. In year 2008, ADASYN Approach for Imbalanced Learning is proposed by Haibo He et al.[9]. They have presented a new adaptive synthetic sampling technique for learning from imbalanced datasets. The necessary idea of this method is to use a weighted distribution for minority class instances according to the level of difficulty in learning, in which more synthetic samples is produce for minority class instances that are difficult to learn as compared to those minority instances that are not difficult to learn. ADASYN technique increase learning performance in two ways: (a) adaptively shift the classification decision boundary toward the hard to learn instances. (b) reducing the bias which is introduced by the class imbalance. They focus on the two-class classification problem for imbalanced data sets, a topic of major focus in recent research activities in the research community. ADASYN is based on the idea of adaptively generating minority data instances according to their distributions: more synthetic samples is generated for minority class instances that are difficult to learn compared to those minority instances that are easier to learn. The ADASYN method can not only reduce the learning bias introduced by the original imbalance data distribution, but can also adaptively shift the decision boundary to focus on those difficult to learn instances. The major objective here is to - reducing the bias and adaptively learning. Based on the original data distribution, ADASYN can adaptively generate synthetic data instances for the minority class to reduce the bias introduced by the imbalanced data distribution. Furthermore, ADASYN can also autonomously shift the classifier decision boundary to be more focused on those difficult to learn instances, therefore improving learning performance. These two objectives are accomplished by a dynamic adjustment of weights and an adaptive learning procedure according to data distributions. In year 2004,Hongyu Guo et al.[23] have proposed DataBoost-IM method which generates the features of the synthetic instances individually. Databoost generates each feature value based on Gaussian distribution within an empirical range [min, max]. In this work, they have described a new technique that combines an ensemble-based learning algorithm, and boosting with data generation to increase the estimation power of classifiers against imbalanced datasets including two classes. In the DataBoost-IM technique, difficult instances from both the classes are identified during execution of the algorithm. Subsequently, the difficult instances are used to separately generate synthetic instances for both the classes. The synthetic samples are then added to the training set, and the class distribution and the weights of the various classes in the new training set are re-balanced. In this work, they discuss a new technique for learning from imbalanced data sets, DataBoost-IM, that combines boosting procedures and data generation to increase the predictive accuracies of both the majority and minority classes, without forgoing one of the two classes. That is, the aim of this approach is to ensure that the resultant predictive accuracies of both classes are high. This approach differs from prior work in the following ways. Firstly, they separately identify hard instances from, and generate synthetic instances for both the classes. Secondly, they generate synthetic instances with bias information toward the hard instances on which the next component classifier in the boosting procedures needs to focus. That is, they provide additional knowledge for the majority as well as the minority classes and thus prevent boosting over- emphasizing the hard instances. Thirdly, the class frequencies in the new training set are rebalanced to make easier the learning algorithm's bias toward the majority class. Rebalancing thus involves the utilization of a reduced number of instances from the majority and minority classes to ensure that both classes are represented during training. Fourthly, the total weights of the various classes in the new training set are re-balanced to force the boosting algorithm to focus on not only the hard instances, but also the minority class instances. In this way, this work focused on improving the predictions of both the minority and majority classes. In recently 2014, Sukarna Barua et al. suggested MWMOTE algorithm for imbalanced Data Set Learning[26]. This work identifies that most of the existing oversampling techniques may generate the wrong synthetic minority instances in some scenarios and make learning tasks difficult. To this end, a novel technique, called Majority Weighted Minority Oversampling method, have presented for efficiently handling imbalanced learning issue. MWMOTE first identifies the hard-to-learn minority class samples and assigns them weights according to their Euclidian distance from the nearest majority class instance. It then generates the synthetic instances from the weighted informative minority class instances using a clustering method. This is done in such a manner that all the generated instances lie inside some minority class cluster.

Some of the most popular approaches to deal with imbalanced learning problems are based on the synthetic oversampling methods [3], [8], [9]. In this work, authors illustrate that in some scenarios many of these methods become inappropriate and fail to generate the useful synthetic minority class instances. In this respect, they propose a new synthetic oversampling method, i.e., Majority Weighted Minority Oversampling Technique (MWMOTE), whose goal is to alleviate the problems of imbalanced learning and generate the useful synthetic minority class instances. The essences of the proposed method are: 1) selection of an appropriate subset of the original minority class instances, 2) assigning weights to the selected instances according to their importance in the data, and 3) using a clustering approach for generating the useful synthetic minority class instances. In 2011, Structure Preserving Oversampling

technique for Imbalanced Time Series Classification have proposed by Hong Cao et al.[24]. This work presented a novel structure preserving oversampling technique for categorizing imbalanced time series data. This method generates synthetic minority instances based on multivariate Gaussian distribution by regularizing the unreliable Eigen spectrum and forecasting the covariance structure of the minority class. By creating variances in the trivial Eigen feature dimensions and maintaining the covariance structure, the synthetic instances spread out effectively into the void region in the data space and it is not closely bind with existing minority class instances. Many real-world learning applications in a wide range of domains, such as entertainment, network security, finance, aerospace, and medicine, involve time series data. A time series instance is an ordered set of real-valued variables that are extracted on a continuous signal, which can be either in the time or spatial domain. Due to its sequential nature, variables that are close in a time series are often highly correlated. One of the best- known learning methods for time series classification is the one nearest neighbor (1NN) classifier with dynamic time warping (DTW). The distance between two instances, known as warping distance, is computed by searching for the optimal mapping path to align the two time series sequences. The classification of a test sample is then based on the top one nearest training neighbor. Imbalanced time series classifications are difficult because of its high data dimensionality and inter-variable correlation. Though oversampling is effective for rebalancing the class balance, but still, it has not been sufficiently explored for imbalanced time series classification due to the complexity of the problem. To achieve the oversampling in general, two existing approaches can be adopted. The first approach interpolates between selected positive samples and their random positive nearest neighbors for generating the synthetic samples. Well-known oversampling methods that adopt this approach are SMOTE [3], Borderline-SMOTE [8] and ADASYN [9] as discussed earlier. The second oversampling approach is to generate the features of the synthetic instances individually. A major method is DataBoost, which generates each feature based on Gaussian distribution within an empirical range [min, max] as seen earlier. These two approaches have been shown to work fairly well for various imbalanced non-time series classification datasets. However, according to author's opinion they may not be enough for oversampling largely imbalanced time series datasets. The adjacent variables in the time series are usually not independent but highly correlated. The random data variances introduced by both traditional oversampling approaches will weaken or even destroy the inherent correlation structures in the original time series data, resulting in non-representative synthetic training instances with excessive noise that confuse the learning. As such, Hong Cao et al. proposed a new structure preserving oversampling method for a binary time series classification task. This method is designed to preserve the covariance structure in the training time series data by operating in the corresponding Eigen spectrum in two subspaces, a reliable subspace and a unreliable subspace, as follows: 1) The synthetic instances are generated by forecasting and maintaining the main covariance structure in the reliable Eigen subspace; 2) A regularization procedure is further employed to understand and fix the unreliable Eigen spectrum. This helps create some buffer variances of the synthetic data in the trivial Eigen subspace to improve the generalization performance on the unseen data. This is the first oversampling technique that preserves the covariance structure in imbalanced learning. In conjunction with Support Vector Machines (SVM), author's show that SPO outperforms other oversampling methods. In recently 2013, Hong Cao et al. suggested Integrated Oversampling (INOS) for Imbalanced Time series Classification [25]. They focused on the problem of Imbalanced learning issue. To address this issue they introduce a new technique of oversampling i.e., Integrated Oversampling technique. They have noted that, the interpolation based approach work well with imbalanced learning issue, but the problem with that is, they are not sufficient for the task of oversampling highly imbalanced time series data sets.

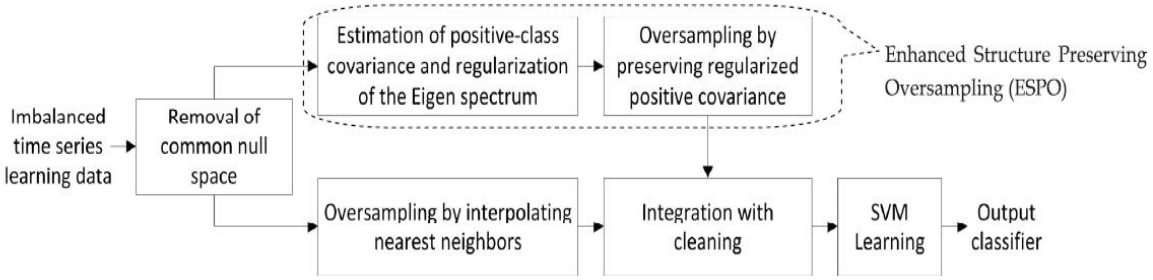

Figure. 1. Block diagram of the integrated oversampling framework.

Hong Cao et al. designed an Integrated Oversampling (INOS) Technique with two objectives in mind: One is to preserve the regularized Eigen covariance structure which can be estimated using the limited positive time series instances, and the other is to be able to provide enough emphasis on the key minority instances with the remaining oversampling capacity. For the first objective, a new enhanced structure preserving oversampling (ESPO) has proposed. ESPO performs oversampling in the transformed signal space in the following steps:

a) Generating the synthetic instances by estimating and maintaining the main covariance structure in the reliable Eigen subspace;
b) Inferring and fixing the unreliable Eigen spectrum, this is insufficiently estimated due to the limited number of positive instances, using a regularization procedure.

For the second objective, they use an interpolation based method to generate a small percentage of synthetic instances so as to emphasize on the border set of existing instances, which are critical for building accurate classifiers in the subsequent steps.For evaluation, INOS with support-vector machines (SVM) classification is used as shown in block diagram of INOS in Figure1. Compared with previous SPO work, the current proposed INOS technique differs and performs better in the following aspects: a) the oversampling is performed in the signal space with improved efficiency and no risk of artificially introducing variances in the common null space.

b) The cleaning mechanism is redesigned to remove the "noise links" or pairs of positive and negative instances on the classification border with good efficiency.

c) A small percentage of oversampling capacity for protective interpolation-based oversampling on the positive-class boundary is reserved, which produces better classification performance.

5. CONCLUSION AND FUTURE WORK

In this paper, the state of the art methodologies to deal with class imbalance learning problem has been reviewed. The imbalanced learning problem for time series classification is much more daunting than typical imbalanced classification problems because of its high dimensionality. Very often, the number of available samples in the minority class is few as compared with the dimensionality. The inherent data complexity of time series classification suggests that it is sensible to address the imbalance problem at the data level using oversampling, as oversampling has been found to be effective for reestablishing the class balance at the data level for generic imbalanced classification problems. To solve this problem the INOS approach performs well for imbalanced time series classification. By examining the characteristics of various time series data sets a meta-learning algorithm can be developed that estimates the best classification methodologies.

REFERENCES

[1] Juanli Hu, Jiabin Deng, Mingxiang Sui, A New Approach for Decision Tree Based on Principal Component Analysis, Proceedings of Conference on Computational Intelligence and Software Engineering, page no:1-4, 2009.

[2] Huimin Zhao and Atish P. Sinha, An Efficient Algorithm for Generating Generalized Decision Forests, IEEE Transactions on Systems, Man, and Cybernetics —Part A : Systems and Humans, VOL. 35, NO. 5, Page no: 287-299, Septmember 2005.

[3] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique,"J. Artificial Intelligence, vol. 16, pp. 321-357, 2002.

[4] A. Estabrooks, T. Jo, and N. Japkowicz, "A Multiple Resampling Method for Learning from Imbalanced Data Sets," Computational Intelligence, vol. 20, pp. 18-36, 2004.

[5] David Hand, HeikkiMannila, and Padhraic Smyth. Principles of Data Mining. MIT Press, August 2001.

[6] Jiawei Han and MichelineKamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, April 2000.

[7] G.E.A.P.A. Batista, R.C. Prati, and M.C. Monard, "A Study of the Behavior Of Several Methods for Balancing Machine Learning Training Data," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 20-29, 2004.

[8] H. Han, W.Y. Wang, and B.H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," Proc. Int'l Conf. Intelligent Computing, pp. 878-887, 2005.

[9] H. He, Y. Bai, E.A. Garcia, and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," Proc. Int'l Conf. Neural Networks, pp. 1322-1328, 2008.

[10] J. Wu, S. C. Brubaker, M. D. Mullin, and J. M. Rehg, "Fast asymmetric learning for cascade face detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 3, pp. 369–382, Mar. 2008.

[11] N. V. Chawla, N. Japkowicz, and A. Kotcz, Eds., Proc. ICML Workshop Learn. Imbalanced Data Sets, 2003.

[12] N. Japkowicz, Ed., Proc. AAAI Workshop Learn. Imbalanced Data Sets, 2000.\

[13] G. M.Weiss, "Mining with rarity: A unifying framework," ACM SIGKDD Explor. Newslett., vol. 6, no. 1, pp. 7 19, Jun. 2004.

[14] N. V. Chawla, N. Japkowicz, and A. Kolcz, Eds., Special Issue Learning Imbalanced Datasets, SIGKDD Explor. Newsl.,vol. 6, no. 1, 2004.

[15] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory Undersampling for Class-Imbalance Learning," IEEE Trans. System, Man and Cybernetics, vol. 39, no. 2, pp. 539-550, Apr. 2009.

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

[17] Y. Sun, M.S. Kamel, A.K.C. Wong, and Y. Wang, "Cost-Sensitive Boosting for Classification of Imbalanced Data," Pattern Recognition, vol. 40, no. 12, pp. 3358-3378, Dec. 2007.

[18] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," Neural Netw., vol. 21, no. 2–3, pp. 427–436, 2008.

[19] A. Freitas, A. Costa-Pereira, and P. Brazdil, "Cost- sensitive decision trees applied to medical data," in Data Warehousing Knowl. Discov. (Lecture Notes Series in Computer Science), I. Song, J. Eder, and T. Nguyen, Eds.,
[20] K.Kilic¸,O¨ zgeUncu and I. B. Tu¨rksen, "Comparison of different strategies of utilizing fuzzy clustering in structure identification," Inf. Sci., vol. 177, no. 23, pp. 5153–5162, 2007.
[21] M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, and R. H. Moss, "A methodological approach to the classification of dermoscopy images," Comput.Med. Imag. Grap., vol. 31, no. 6, pp. 362–373, 2007.
[22] X. Peng and I. King, "Robust BMPM training based on second-order cone programming and its application in medical diagnosis," Neural Netw., vol. 21, no. 2–3, pp. 450–457, 2008.Berlin/Heidelberg, Germany: Springer, 2007, vol. 4654, pp. 303–312.
[23] H. Guo and H.L. Viktor, "Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost-IM Approach," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 30-39, 2004.
[24] H. Cao, X.-L. Li, Y.-K. Woon, and S.-K. Ng, "SPO: Structure Preserving Oversampling for Imbalanced Time Series Classification," Proc. Int'l Conf. Data Mining (ICDM '11), pp. 1008-1013, 2011.
[25] Hong Cao, Xiao-Li Li, David Yew-Kwong Woon, and See-Kiong Ng," Integrated Oversampling for Imbalanced Time Series Classification," IEEE ," IEEE Trans. Knowledge and Data Eng., VOL. 25, NO. 12, pp.2809-2822, 2013
[26] Sukarna Barua, Md. Monirul Islam, Xin Yao, Fellow, IEEE, and Kazuyuki Murase," MWMOTE—Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning," IEEE Trans. Knowledge and Data Eng., VOL. 26, NO. 2, pp.405-425, 2014

AUTHOR'S

Reshma K. Dhurjad received the B.E degree in Computer Engineering from K.K.Wagh Institute of Engineering Education & Research, Nasik, Savitribai Phule Pune University in 2012. Now pursuing M.E. from K.K.Wagh Institute of Engineering Education & Research, Nasik, India.

Prof.S.S.Banait, Assistant Professor, Department of Computer Engineering, K. K. Wagh Institute of Engineering Education & Research, Nasik, India.