IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Vol.-6, Issue-12, Dec 2018 E-ISSN: 2347-2693

A Survey of Different Techniques to Handle An Unbalanced Dataset

Pooja Yerawar [1*], Ganesh Pakle[1]

[1]Department of Information Technology, SGGS IE&T, Vishnupuri, Nanded, India

[*]*Corresponding Author: yrw.puja22@gmail.com, Tel.: +91-9980498957*

Available online at www.ijcseonline.org

Accepted: 22/Dec/2018, Published: 31/Dec/2018

Abstract— Researchers has a big challenge to handle the unbalanced data, which is an issue found in many real-world applications in engineering. Dataset is unbalanced means at least one class has very fewer examples than another class. In such dataset, examples are taken as majority class (i.e. negative) and minority class (i.e. positive). This paper contains a survey of what is mean by imbalance data, an issue with it, its challenges, examples of applications, different approaches to rebalance the data like ensemble techniques(like boosting, bagging), sampling, feature selection, algorithmic to increase the performance of classification have been proposed.

Keywords— Imbalanced data, classifiers, sampling, feature selection, ensemble methods, hybrid method.

I. INTRODUCTION

Unbalanced problems mainly arise due to a number of the examples in one set is generally much smaller (greater) than that of the examples outside the set. These are a particular case of data set for classification issue where the classes are noneven. There are two different classes: The majority class and the minority class i.e. negative and positive class respectively. Suppose these types of data is a new challenging issue for Data Mining and machine learning since normally standard classifiers consider a balanced training set. Classification of data sets done by different classifiers according to its class labels. This problem is common and can be seen in various real-world disciplines like fraud detection, medical diagnosis, oil spillage detection, facial recognition, anomaly detection, cultural modeling, fault detection, text categorization, and satellite images [1]. In two classes proportion between the majority (negative)and minority (positive) class may be 90:1,100:1 and 800:1; it means, the examples of majority class are more than minority class examples.

In high dimensional data, class unbalance issue is very complex. In unbalanced datasets, the class proportion is sufficiently critical that classifier made biased with a few classes (i.e. majority class). Performance bias implies arrangements which give high exactness on the negative classes and less precision on the minority classes. Non even spreading of class samples can decrease the implementation of different classifiers by recent studies. The exact classification of instances from minority (positive) class can be essential than that of exact classification of instances from the majority class. This paper contains, different issue of data unbalance in classification, the productive measures are given by different authors to manage data unbalance is shown and different techniques to handle data unbalance issue is differentiated among them [2].

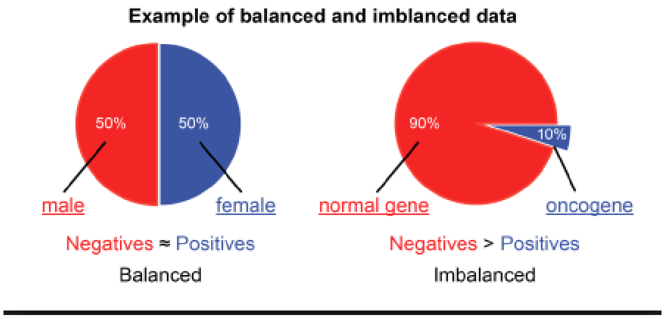


Fig 1.Example of Balance and imbalanced data

Before this, there were many solutions given for class unbalance issue, like the data level, algorithmic levels, and sampling (Oversampling and undersampling) techniques. Data level [3], includes numerous types of resampling techniques alike random undersampling, oversampling, random oversampling with replacement and combinations of all methods. Algorithmic level mainly contains, adjust the weights of the different classes to solve the class unbalance, at the tree leaf by adjusting the probabilistic estimate (in decision trees), by adjusting the decision threshold, and recognition-based (i.e. learning from one class) rather than discrimination-based (two class) learning. After merging various methods these are used to handle unbalance class issues. It is the issue in binary class as well as in multi-class. Till now many of solutions are given for binary class unbalance issue, but the issues concerned to the multi-class unbalance are not getting any solution. This paper describes many methods for handling unbalance dataset issues.

Many techniques have been defined to solve the unbalanced issues. It has three groups which include: (1) Rebalancing the distributions of class (2) Adjust the classifiers to the unbalanced datasets by considering the cost or weight for the not correctly classified examples, and (3) ensemble learning technique. Re-sampling techniques like over-sampling and under-sampling are famous techniques because of its simple execution and comparatively well performance. Additionally, in general, the oversampling method outperforms than the under-sampling method is shown. Hence, many of the famous methods which deal with unbalanced learning issues depend on the oversampling method.

By research it was found that there were two types of unbalance data one is Binary class data unbalance and another is multi-class data unbalance.

1) Binary class data unbalance

Binary dataset means it has only two classes of a dataset. Binary class data unbalance issue means if in the binary dataset there exists a class which is shown by only a few numbers of examples. In binary class dataset to separate two classes, zero class thresholds are generally used so there is no need to recognize the boundaries of classes in a dataset.

2) Multiclass data unbalance

Multiclass data means which contains more than two classes. Data unbalance issue create additional overheads in a multi-class dataset. In the multi-class dataset, simple and efficient zero class thresholds cannot be used. Some methods like Complex Static Search Selection or Dynamic Search Selection require to be used. Sometimes to classify a dataset, the multiclass issue is required to be divided into many binary class issues [2].

Rest of the paper is organized as follows, Section I contains the introduction of what is unbalanced data, binary and multiclass data, section 2 covers different issues and examples with an unbalanced dataset. Section 3 covers background work to solve unbalance problem with different approaches like feature selection, algorithmic level, ensemble, and sampling approaches. Lastly, in section 4 the concluding remark is given.

2. Issues and examples with an unbalanced dataset

Issues which occur when a dataset is unbalanced are:

- The weight of missing a negative class is much lesser than that of missing a positive class (i.e. its cost is higher).
- Most learning frameworks are not set up to adapt to an expansive contrast between the many of cases having a place with each class.
- When data is unequal, algorithms for classification under-performs.

The unbalance issue is a respective problem, which based on:

1) The proportion of the negative to positive examples (ratio).
2) The difficulty of the idea shown by the data.
3) The General size of the training set.
4) Involvement of classifier.

Examples:

1) Detection of frauds:

Transactions whichever done online fraud detection is an issue of significant economic effect. The quantity of fraudulent transactions is normally a little part of whole transactions and thus this issue is frequently referred to as average data unbalance issue. In many time, a system which is used for a fraud detection will potentially fraudulent transactions to be assessed manually by an expert.

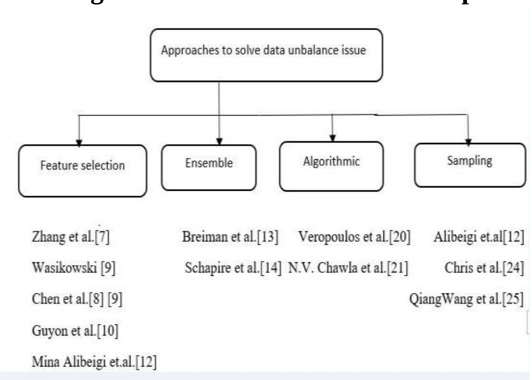
2) Categorization of Products:

Online business retailers sort their item list into practical gatherings to help seek recovery. There is substantial variation in the number of things having a place with every classification. For example, there are just a couple of Samsung mobile models while the quantity of Samsung mobile wearable (e.g. charges, cases, styli, and so forth) is a few hundred folds more. There will undoubtedly be a significant measure of cover in the representation and pictures of things from these two classes. A programmed item arrangement framework can possible to jumble between the two classes. On the off chance that the retailer is enhancing for income, it will be good to guarantee all mobiles are classified effectively at the fear of characterizing a couple of Samsung wearable as Samsung mobiles. Data of the product is different and unbalanced. A single product can have a whole set of attribute data, it may be entirely missing in other product (e.g. color, a fat content of the milk or volume of milk). Considering all accessible item factors would lead a blast of missing qualities, which makes model convergence that substantially harder. To deal with this issue, we chose to keep it straightforward and just utilize item names, pictures, and explanation as our prediction factors, since they are accessible for the greater part of items and convey the most valuable data.

3) Diagnosis of disease:

The fraction of sound individuals outnumber those acted with it, for any given disease. If there should arise an occurrence of uncommon alignment, it is a repetition to state that the dataset is profoundly unbalanced. On the off chance that a computerized classification framework is utilized to predict the presence of the sickness (likely pursued by a specialist evaluation), it is very valuable to have a review on the disease class to be as near 1 as could be expected under the circumstances. In this specific case, on the minority class high exactness is important since a significant measure of master examination might be required for staying away from false positive disease prediction on sound individuals [4].

3. Background work to solve unbalance problem



1. Feature selection approaches:

This section includes a few of the feature selection methods for unbalanced datasets. This method selects important features to increase the performance and exactness of the classifier. Due to non-relevant attributes in unbalanced dataset performance of the classifier may be lower. Its metrics can be spitted into two i.e. one-sided takes only minority features [9] while two-sided merge the features from minority and majority [10].

Zhang et al. [5] defined framework of two-sided choices of features on unbalanced data, it merges definitely the minority and majority features in optimal fashion approximately. Before this Odds ratio (OR) and Correlation coefficient (CC) are one-sided metrics although chi-square (CHI) and information gain (IG) are two-sided. The authors give easy ideas of existing measures transformation so they consider attributes for the majority and minority classes independently.

In the [11] last decade, the unbalance class issue is generally followed by the problem of high dimensionality and little sample size of the dataset [6]. A few particular examples involve but they are not restricted to data analysis of gene expression (mass spectrometry and microarray data), face recognition, text mining, and fraud detection [12]. A little sample size issue can create a classifier which not only discovers aspects of the data precisely but also over-fit the training data and it produces false predictions [6] on test data. Usually, the little sample size has been measured broadly in whatever work done before. Algorithms like Dimensionality reduction such as principal component analysis (PCA) and add-ons of it have an answer to this issue [12] because of the certainty which is a good option of a process to expand generality of probably a classifier is a feature selection [13]. To deal with unbalanced datasets, integration of little sample size and unbalanced data is a current solution which is considered in [12] to a little degree of fact. There are various ways used to gear the class unbalances issue [6]. To battle the high dimensional unbalance issue approaches like sampling and algorithmic may not be adequate.

Wasikowski and Chen et al. [14,6]defined in unbalanced data sets which are high dimensional, feature selection can only battle the unbalance class issue; still, in his work, Elkan got that feature ordering procedure is not adequate to gear this issue and the co-operation among various features need to examine in the selection procedure of features. With this, he also listed that limitation of many of the feature selection methods which did not examine highly corresponding features as they were convicted to be unwanted.

With this Guyon [7] proposed a powerful theoretical analysis of the limits of feature ordering procedures and she told that those features which are not in use (non-relevant) by themselves, may be helpful in association with alternative features [7]. The execution period for catching the good subset of the feature among a desirable subset of the feature is having an order O(2n), in this n is the no of features of the data set when this is used for high dimensional data sets the execution period is unmanageable. Besides, subset of feature selection techniques such as embedded ones and wrappers deal with the coordination among features in the subset selection phase, can catch the feature subset which overfits the training data [6]; though feature ordering processes do not go through these issues to deal with high dimensional data sets [6] and when feature ordering processes are not more advantageous, they may be referred as a result of their linear execution period in the size of features of the data sets [6].

There are a few of the feature selection methods for unbalanced datasets. Xue-wen Chen and Michael Wasikowski et al. proposed FAST it depends on the area under the receiver operating characteristics (ROC) which are created by moving the decision boundary of only one feature classifier with thresholds set using an even-bin distribution. Between the mean of the two classes, there is midpoint many of the single feature classifiers put decision boundary. We cannot say this is a prime choice for the decision boundary. This issue can be solved by classification of the instances using many thresholds and collecting statistics about the performance at every boundary. We can formulate a ROC curve and measure the AUC, after calculating the TP rate and FP rate at every threshold. Because especially for unbalanced data classification issues the area under the receiver operating characteristics is a powerful predictor of performance, this result can be used for listing the features and those features are selected with the highest AUC since they have the best predictive ability for the dataset. It uses ROC curves to list the features which create the new problem, for deciding where to put the thresholds. To solve both of these issues this method utilizes an even bin distribution. Other than choosing the bin width and changing the number of points in every bin, they choose the no of points to fall in every bin also change the bin width. So, both of the above goals are fulfilled. The regions in the feature space that have lesser instances are covered by broader bins, and regions that have many instances are covered by smaller bins. Afterward, this method takes the mean of every instance in every bin as the threshold and classifies every instance depending on this threshold [15].

In 2011, Mina Alibeigi et.al. given Unsupervised Feature Selection method based on the Distribution of Features Attributed to unbalanced Data Sets [8] which discard unnecessary features from the authentic feature space based on the distribution of features. All the features are first scaled in the range [0, 1] in this method. After that, the probability density function (PDF) of every feature is predicted which gives a good analysis of the distribution of instances for a specific feature. Next to many a time the PDF of one feature is compared with PDF of other remaining features are measured. Finally, features which have a higher counter of being identical to other features are eliminated.

2. Ensemble approaches

By combining several models ensemble learning helps to progress machine learning results. Mainly ensemble methods are used to increase the prediction exactness. Boosting is the most popular ensemble approach that re-samples instances adaptively as indicated by their costs also creates a highly exact ensemble of classifiers whose single classifier has moderate precision. The cost of wrongly classified instances by previous classifiers are balanced progressively so every one of these cases can be fixed more on by the classifiers. These techniques are directly applied to increase the classification exactness of the unbalanced datasets. The diversity of the single classifier is a very important aspect in ensemble techniques which finds the last prediction exactness of an instance. To describe and expand diversity is a very important task. This methodology permits the creation of better predictive presentation compared to a single model. Ensemble approaches are meta-algorithms that merge numerous machine learning techniques into one predictive model in order to decrease variance (bagging), improve predictions (stacking) or bias (boosting).

Usually, ensemble techniques use a single base learning algorithm to create same base learners, i.e. leading to homogeneous ensembles, learners of the same type. With this, there are also some methods that use heterogeneous learners, i.e. leading to heterogeneous ensembles, learners of different types. For ensemble methods to be more exact than any of its single members, the base learners have to be as accurate as possible.

Breiman et al. [breiman1996bagging] defined bootstrap aggregation, or bagging. This process can be used with many classification procedures and they applied regression approaches to decrease the variance related to a prediction which enhances the prediction method. Prediction technique is used to every bootstrap example after that its results are grouped by taking an average for the regression and by taking voting for classification to get the general prediction. By doing analysis on actual and assumed data sets using classification and regression trees and sub-part selection in linear regression presents bagging can provide considerable increases in exactness. The component is the fluctuations of the prediction technique. Bagging can raise exactness if changing the learning set may create appropriate variations in the predictor built.

Yoav Freund et al. [17] in 1996 introduced new "boosting algorithm i.e AdaBoost" it may automatically decrease the error of any learning technique that continuously creates classifiers whose performance is somewhat more than random guessing with this they defined the related concept of a “pseudo-loss” it is a process for forcing a learning technique of multi-label concepts to focus on the labels which are very hard to distinguish. They also introduced different experiments that carried out to evaluate how nicely AdaBoost work on real learning issues with and without pseudo-loss. They evaluate two parts of experiments. The first part differentiated with boosting to Breiman’s “bagging” process when this is used to combine different classifiers (containing decision trees and single attribute value tests). They have deliberated in much detail the performance of boosting by using a nearest-neighbor classifier on an OCR issue in another set of experiments.

Schapire et al. in 1990 [18] were defined Boosting which is also called adaptive sampling and ARCing. Schapire demonstrated that a weak learner may be changed into a strong learner in the form of probably approximately correct (PAC) learning basis. The greatest demonstrative algorithm in this category is AdaBoost [17], this has been selected as the top ten data mining procedures or techniques [19] which was the first most appropriate approach to Boosting. AdaBoost is mainly acknowledged to decrease bias (rather than variance) [20], and likewise, support vector machines (SVMs) which boosts the borders [21]. It uses the complete data-set to train every classifier in sequence, but next to round, it gives more focus to tough the samples, with the aim of correctly classifying examples in the next round that was not correctly classified throughout the existing iteration.

Thus, it concentrates more on those samples which are difficult for classification, the weight is counted by the quantity of focus, which is primarily equivalent for all examples. After every iteration, the costs of wrongly classified examples are enlarged; contradictory to this, the weights of correctly classified examples are reduced. In addition to this, one more weight is allocated to every single classifier which depends on it's all over exactness that is used in the test phase afterword's; more confidence is given to more accurate classifiers. Finally, by majority voting of every classifier, the class label is selected when a new example is submitted.

3. Algorithm approaches

This approach is mainly defined to resolve the class unbalance issue by creating or modifying the already available classification techniques [dubey2014analysis]. In the learning method, Cost-sensitive approaches include dissimilar misclassification costs for every class.

Veropoulos et al.[veropoulos1999controlling] defined a biased support vector machine (B-SVM)technique through the modification of objective function after setting dissimilar cost to the negative and positive of the unbalanced data, that makes the learned hyperplane far from the minority class. They defined two approaches to adjust the specificity and sensitivity of svms and their performances using ROC. Thus, this process did not examine the dissimilar involvement of the instances in the similar class when setting decision hyperplane, so this method is not so useful to increase the prediction exactness of the minority.

Nitesh V. Chawla et al.[chawla2008automatically] defined cases which found infrequently like instances of disease, regions of interest in large-scale simulations and fraud which requires high cost relatively for the misclassification of infrequent measures. To make models with high minority class precision the informational collection is regularly re-examined. Thus, this technique faces typical things like how to decide naturally the adequate amount and form of sampling? To solve this issue, they defined a wrapper model which finds the quantity of re-sampling for a data set that is based on optimizing evaluation functions like the Area Under the ROC Curve (AUROC), cost, cost-curves, f-measure, and the cost dependent f-measure. Their study of the wrapper is twofold. 1. They found the communication between various calculation and wrapper optimization functions.2. They give an arrangement of results in a cost-sensitive condition. After comparison of the performance of the proposed method versus cost-sensitive learning methods like MetaCost and the Cost-Sensitive Classifiers, they found the proposed method outperforms than cost-sensitive classifiers.

4. Sampling approaches

An easy way to balance the unbalanced dataset is sampling the data. The sampling techniques are used to re-balance the distribution of data to decrease the outcome of the unbalanced or skewed class distribution in the learning process [27, 28]. The simple oversampling and under-sampling techniques are random over-sampling and random under-sampling respectively, which duplicates or removes randomly the samples of the minority or majority. The result of random sampling techniques is the development of the classification performance of an unbalanced dataset; then also there could some issues which arise.

Sampling techniques are divided into three groups as follows:

1) Oversampling

This is one of the sampling techniques which balance the data set by copying the instances of a minority class. It is also called as sampling. The main benefit of this technique is that there is no any loss of data. The drawback of this method is that it may cause overfitting and it can introduce an added computational overhead. It is also divided into two forms. Random Oversampling and Informative Oversampling. Random Oversampling is the technique which balances the class distribution by duplicating the randomly selected minority (Positive) class instances. Informative Oversampling technique synthetically creates minority (negative) class instances based on a pre-specified condition [2]. Random oversampling is one of the simple and effective methods of resembling. In this, we select members from the positive class randomly; then these randomly selected members are then replicated and added to the new training set. In random oversampling we must need to remember two things: First, it chooses examples randomly from the original training set, not the new training set. Second, it always oversamples randomly with the replacement. If without replacement we were to randomly oversample before we got the chosen balance between the majority and minority class we would reduce all members of the minority class. The no of Oversampling methods are available in the literature like SMOTE, Borderline SMOTE, OSSLDDD-SMOTE etc.

2) Undersampling

This is another effective technique for balancing data. This method trains the classifier by using a subset of the majority class. Opposite to oversampling in undersampling, we delete some samples of the majority class. Undersampling techniques also divided into random and informative. Random Undersampling is very simple, this method balance class distribution by random elimination of majority class instances. The other is Informative Undersampling, this technique chooses only the required majority class examples which are based on a pre-specified selection measure to make the dataset balanced. Informative Undersampling may be passive or active. Passive selection approaches are defined as a pre-processing method for choosing informative examples for a classifier. Informative examples are inquired throughout the construction procedure of the classifier in Active selection approaches. The most commonly used pre-processing method is random majority undersampling (RUS). In random undersampling, Examples of the majority class are randomly rejected from the dataset.

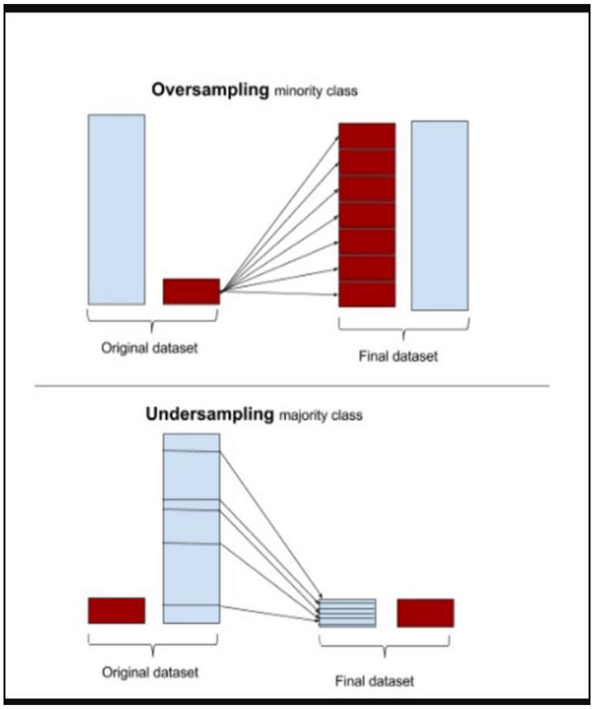


Fig 2.Oversampling and Undersampling

Thus, the main disadvantage of under-sampling is that possibly useful information is ignored. To increase the performance of random sampling there are many ways, such as Tomek links, Condensed Nearest Neighbour Rule and One-sided selection etc.

3) Hybrid Methods

Hybrid sampling methods combine oversampling and undersampling approaches [8] to address the class unbalance issue or it combines the strengths of any two approaches to balance the data.

Chris et al. [25] defined a hybrid approach to resolve data unbalance issue that is called RUSBoost. This method merges sampling and boosting methods to solve the data unbalance issue. A drawback of RUSBoost is that it neither contains other learners nor considers performance metric.

Qiang Wang et al. in 2014 presented [26] a hybrid sampling SVM method it also combines an oversampling method and an under-sampling method to address the classification issue of unbalanced data. The proposed method firstly uses an under-sampling method to remove a few examples of the majority class with fewer classification data and after that, they use an oversampling method to progressively create a few new positive examples. So, to change an original unbalanced training dataset, a balanced training dataset is produced. By using some enhanced technique we can improve our current model and have some future work [29] [30].

After comparison of all our approaches, we know that the ensemble and hybrid methods outperform than others because it combines the strength of different classifiers.

IV.CONCLUSION

Data unbalance is a big issue for researchers in many real-world applications. These are a particular case of data set for classification issue where the classes are non-even. There were some re-sampling methods to increase classification performance on minority class when data unbalance is present. In this paper, work done to solve the class unbalances issue has been reviewed. First this paper we give the theoretical idea of data unbalance, its issues and examples then discuss different approaches to avoid and handle data unbalance in classification afterward we compare different methods to handle data unbalance issue. Because it combines the strength of different classifiers we get that ensemble and hybrid methods perform well on unbalanced data after comparison of different methods.

REFERENCES

[1] Sonak and R. A. Patankar, “A Survey on Methods to Handle Imbalance Dataset,” *International Journal of Computer Science and Mobile Computing*, vol. 4, no. 11, pp. 338–343, 2015. [Online].Available:http://ijcsmc.com/docs/papers/November2015 / V4I11201573.pdf

[2] Singh and A. Purohit, “A survey on methods for solving data imbalance problem for classification,” *International Journal of Computer Applications*, vol. 127, no. 15, pp. 37–41, 2015.

[3] N.Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *Jair*, vol. 16, pp.321–357, 2002.

[4] More, “Survey of resampling techniques for improving classification performance in unbalanced datasets,” vol. 10000, pp.1–7, 2016. [Online]. Available: http://arxiv.org/abs/1608.06048

[5] Z. Zheng, X. Wu, and R. Srihari, “Feature selection for text categorization on imbalanced data,” *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 80–89, 2004.

[6] M. Wasikowski and X. W. Chen, “Combating the small sample class imbalance problem using feature selection,” *IEEE Transactions*

[7] Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *Journal of machine learning research,* vol. 3, no.Mar, pp. 1157–1182, 2003.

[8] M. Alibeigi, S. Hashemi, and A. Hamzeh, “Dbfs: An effective density based feature selection scheme for small sample size and high dimensional imbalanced data sets,” *Data & Knowledge Engineering,* vol. 81, pp. 67–103, 2012.

[9] F. Sebastiani, “Machine learning in automated text categorization,” *ACM computing surveys (CSUR),* vol. 34, no. 1, pp. 1–47,2002.

[10] M. Alibeigi, S. Hashemi, and A. Hamzeh, “DBFS: An effective Density Based Feature Selection scheme for small sample size and high dimensional imbalanced data sets,” *Data and Knowledge Engineering,* vol. 81-82, pp. 67–103, 2012. [Online].Available: http://dx.doi.org/10.1016/j.datak.2012.08.001

[11] N. V. Chawla, N. Japkowicz, and A. Kotcz, “Special issue on learning from imbalanced data sets,” *ACM Sigkdd Explorations Newsletter,* vol. 6, no. 1, pp. 1–6, 2004.

[12] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Transactions on Knowledge & Data Engineering,* no. 9, pp.1263–1284, 2008.

[13] X.-w. Chen and M. Wasikowski, “Fast: a roc-based feature selection metric for small samples and imbalanced data classification problems,” *in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM,2008, pp. 124–132.

[14] G. Forman, “An extensive empirical study of feature selection metrics for text classification,” *Journal of machine learning research,* vol. 3, no. Mar, pp. 1289–1305, 2003.

[15] H. Pant and R. Srivastava, “A Survey on Feature Selection Methods For Imbalanced Datasets,” *International Journal of Computer Engineering and Applications,* vol. 9, no. 2, pp. 197–204, 2015.

[16] L. Breiman, "Bagging Predictors," *Machine learning,* vol. 24, no. 2, pp. 123–140, 1996.

[17] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of computer and system sciences,* vol. 55, no. 1, pp. 119–139, 1997.

[18] R. E. Schapire, “The strength of weak learnability,” *Machine learning,* vol. 5, no. 2, pp. 197–227, 1990.

[19] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip et al., “Top 10 algorithms in data mining,” *Knowledge and information systems,* vol. 14, no. 1, pp. 1–37, 2008.

[20] J. Friedman, T. Hastie, R. Tibshirani et al., “Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors),” *The annals of statistics,* vol. 28, no. 2, pp. 337–407, 2000.

[21] Rudin, I. Daubechies, and R. E. Schapire, "The dynamics of AdaBoost: Cyclic behavior and convergence of margins," *Journal of Machine Learning Research,* vol. 5, no. Dec, pp. 1557–1595, 2004.

[22] K. Veropoulos, C. Campbell, N. Cristianini et al., “Controlling the sensitivity of support vector machines,*” in Proceedings of the international joint conference on AI,* vol. 55, 1999, p. 60.

[23] N. V. Chawla, D. A. Cieslak, L. O. Hall, and A. Joshi, "Automatically countering imbalance and its empirical relationship to cost," *Data Mining and Knowledge Discovery,* vol. 17, no. 2, pp. 225–252, 2008.

[24] N. V. Chawla, K.W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique,*” Journal of artificial intelligence research,* vol. 16, pp. 321–357, 2002.

[25] Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, “Rusboost: A hybrid approach to alleviating class imbalance,”

[26] Q. Wang, “A hybrid sampling SVM approach to imbalanced data classification," *in Abstract and Applied Analysis,* vol. 2014.Hindawi, 2014.

[27] G. E. 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.

[28] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* vol. 42, no. 4, pp. 463–484, 2012.

[29] D.K. Mittal, V. Verma, R. Rastogi, "*A Comparative Study and New Model for Smart Mirror*", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.6, pp.58-61, 2017

[30] Dharmendra Sharma and Suresh Jain, "*Evaluation of Stemming and Stop Word Techniques on Text Classification Problem*", International Journal of Scientific Research in Computer Science and Engineering, Vol.3, Issue.2, pp.1-4, 2015

Authors Profile

Ms.Pooja Yerawar pursued Bachelor of Engineering degree in Information Technology from SRPCE Nagpur in year 2012, which is affiliated to Nagpur University. She is currently pursuing Master of Technology (Full time) in Information Technology from Shri Guru Gobind Singhji Institute of Engineering and Technology Nanded which is affiliated to SRT Marathwada University and currently working as a Teaching Assistant in the Department. Her research interest includes Machine learning, Data mining and Data analysis.

Mr. G K Pakle pursued Bachelor of Computer Science and Engineering from Dr. B.A.M. University, Aurangabad, India in 2002 and Master of Engineering from Swami Ramanad Teerth Marathwada University, Nanded, India in year 2011. He is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Information Technology at Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded since 2004. He is a life member of the Computer Society of India and ISTE. He has published more than 10 research papers in reputed international journals including Springer and conferences including IEEE and it’s also available online. His main research work focuses on caching and forwarding in Named Data Networking, Architecture and Design of Future Internet. He has 15 years of teaching experience and 4 years of Research Experience.