K-means and Particle Swarm Optimization based under Sampling Method for Imbalance Datasets

Mohd Ashraf[1], Rajesh Mishra[2]
[1]*Maulana Azad National Urdu University, Hyderabad*
[2]*School of Information & Communication Technology, Gautam Buddha University, Greater Noida*
(E-mail: ashraf.saifee@gmail.com)

Abstract**—**The Imbalanced class problem is a recent challenge in data mining. A dataset is said to be imbalance when their classification categories are not properly defined, and the class which has a few instances as compare to other classes is of more interest from the point of view of the learning task. In recent years, various methods have been proposed for finding a relatively balanced class distribution and equal misclassification costs. This paper provides grouping of two methods. First one is called Supervised and another is called unsupervised learning. Each of these is having its own pros and cons. That's the reason to include both kinds of the method for the proposed solution. Here, there are two kinds of solution for handling the issues of imbalance dataset. These are divided into two categories. The first is called super sampling while another is called under sampling. To do either of these operations, the requirement is of coupling of grouping. Coupling of this grouping is becoming a very critical topic to handle it. This happens just because single grouping does not significantly enhance the performance, especially when the dataset critically suffers from imbalance property. The main objective of this article is to offer an experimental analysis of imbalanced datasets to understand the issue and its solution. The proposed work is implemented in MATLAB 2014. This work shows the performance of coupling of clustering and classification method in order to handle the issue of data imbalance.

Keywords**—***Imbalance dataset, K-means, Particle Swarm Optimization, Sampling methods*

I. INTRODUCTION

Classification is an important task in data mining. Classification algorithm is used to train the model to predict the class level of unseen data. The various classification algorithms such as decision tree, Bayesian network, neural network, nearest neighbor and support vector machines (SVM) have been used to predict the class of unknown data. But all the existing classification methods assume a relatively balanced class distribution [1].

In class imbalance problem the number of instances of one class is much more than the other classes and the class which is of more interest (Minority class or Positive) has few examples as compare to negative (Majority class) examples. When a model prepared with imbalanced data set, it ultimately gives its inclination towards majority class, because they are designed to maximize the overall prediction accuracy. Therefore, standard classifiers ignore overall minority class examples (treating them as noise or outliner) and lose its classification ability in class imbalance problem. For example, in a data set whose imbalance ratio (IR) is 1:100 (i.e., for each example of the minority class, there are 100 majority class examples). A standard classifier may obtain an accuracy of 99% by the ignorance of minority examples, with the classification of all examples as majority. An accurate classification model is one that can provide a higher identification rate of rare examples. Therefore, the class imbalance problem is also referred as a rare class problem.

In recent years, the imbalance data learning issues attend much interest from industries, academics and research teams; refer as the top most challenging issues in the field of data mining [2]. These issues have been observed in several fields like as social sciences , credit card fraud detection , tax payment , customer retention , customer churn prediction , segmentation , medical diagnostic imaging , detection of oil spills from satellite images, environmental studies , bioinformatics and for increasing the value of mammography examinations for the detection of cancer [3]. The importance of the class imbalance problem and its presence in practical applications in the field such as Machine learning and data mining has attracted a lot of research interests. This concept is mostly required in the real world applications, where it becomes expensive for not classifying the examples of the minority class, like as searching of the fraudulent telephone calls, diagnosis of rare diseases, information retrieval, text categorization and filtering tasks [4]. To address this problem various techniques have been developed. These techniques can be categorized into three groups: (1) the nature of the problem (i.e., the kind of data (domain), data complexity, such as overlapping, lack of data and small disjuncts); (2) the possible solution that can predict/identify the class level of unseen data; (3) the appropriate evaluation metrics to measure the classification performance. Within these suggested groups, the most challenging issue is the second one. The possible solution of the class imbalance problem can be divided into two categories as data level and algorithm level solutions. At the data level, the objective is to preprocess the data in advance to remove the effect of skewed class distribution. This is also known as external work process. At the algorithm level, the objective is to generate an efficient algorithm or replace the existing one that can bias toward the positive class [5]. This is also known as internal process. Both the approaches have some drawbacks, some of them are: (1) the data level approach has the drawback of losing some valuable information when majority class samples are under sampled and over-fitting /overgeneralization when minority class samples are oversampled; (2) the drawback of algorithm-based approach is that it requires algorithm specific modification.

Here figure 1, describes the distribution of majority class, minority class and noise samples.

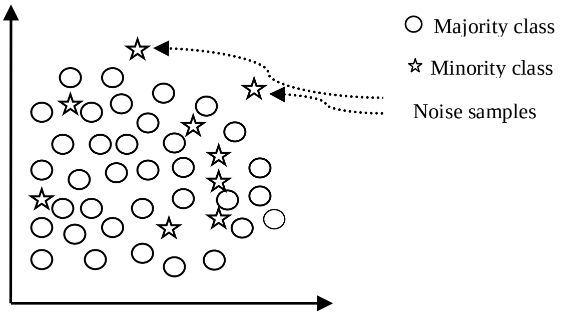

Figure 1: A data set with a between-class imbalance.

Due to lack of unified framework, we need additional research efforts for the advancement of class imbalance problem. The objective of this paper is to review the state of the art techniques to address a two-class imbalance dataset problem and to propose a generalized framework that can be fitted in all types of class imbalance problem.

The rest of this paper is organized as follows: First, section 2 presents the application domains of class imbalance problem. Section 3 describes the nature of the problem and also provides the foundation for our review of class imbalance learning. In section 4, we present the state-of-the-art solutions for imbalanced learning which include sampling methods, cost-sensitive learning methods and various ensemble methods. In section 4, we discuss appropriate measures for evaluating classification performance in the presence of a class imbalance problem. In section 5, we present opportunities and challenges for future research in the field and make concluding remarks.

II. DOMAIN OF CLASS IMBALANCE PROBLEM

The class imbalance problem exists in a large number of domains of importance in the data mining community. The following examples briefly illustrate each one:

A. Medical Diagnosis

Medical databases store huge amount of information about patients and their medical history. The data mining techniques applied on these datasets are used to discover the progression and features of certain diseases. This knowledge can be used for early identification of diseases. But in the medical domain, disease cases are very rare as compared with normal cases and the cost of misclassifying (misdiagnosis) will be fatal as potentially affected patients will be considered healthy.

B. Fraud Detection

Fraud detection in banking transaction, such as credit card fraud is a costly affair in many business organizations. Frauds are detected by analyzing the unusual patterns in transactional databases. But usually in transaction collections, there are many more legitimate users than fraudulent transaction. Therefore, it is difficult to find fraud due to rare cases of fraud transactions.

C. Fault Diagnosis

Due to network-based computer systems, attacks on computer and networks grow rapidly. Therefore, early detection approach used to automate and simplify the manual development of fault diagnosis.

D. Detection of Oil spills

There is only 5% to 10% of oil spills from natural sources. While, most of the pollution caused by ships that want to dispose their waste in the sea. A satellite images based system could be an effective to find illegal dumping and could have significant environmental impact.

The medical diagnosis problem, fraud detection problem and intrusion detection problem are also recognized as an anomaly detection problem. As anomalies are rare as compared with normal observations, the class imbalance problem is thus intrinsic to such kind of applications.

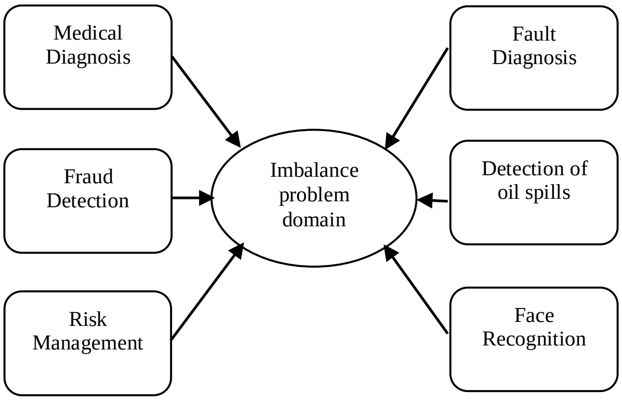

Figure 2: The area which suffers most due to a class imbalance problem.

III. CLASS IMBALANCE PROBLEM

In this section, we first introduce the problem of imbalanced datasets in classification. Then, we present the evaluation metrics for classification problem and then we present several techniques to address the class imbalance problem.

A. Nature of the problem

Define A dataset that has skewed data distribution between its classes can be considered imbalanced. Furthermore, the class which has fewer numbers of instances is usually the class of interest from the point of learning task [6]. This form of imbalance is also referred as a between-class imbalance; e.g. on the order of 100:1, 1,000:1, and 10,000:1. This problem attracts many research interests from researchers due to many real-world classification problems, such as risk management, pollution detection, remote-sensing, fraud detection and medical diagnosis [7]

The implications of imbalanced problem can be highlighted with an example from the medical domain. Consider the example of "Mammography data set", which is used as detection of breast cancer, through detection of characteristic masses. By analyzing the mammography images, collected from a set of distinct patients, the classes that shows "Positive" or "Negative" for an image representative of a "cancerous" or "non cancerous" patient, respectively. In the real world, the non cancerous patients greatly exceed to the number of cancerous patients, i.e., the "negative" samples outweigh over the "positive" samples. Therefore, a classifier is needed that provide a balanced degree of prediction accuracy for both the small and prevalent classes on the data set. But in reality, the standard classifiers have a bias towards the prevalent class having accuracies close to 100 percent and the small class having accuracies of 0 to 10 percent. The standard classifiers overall ignores the small (minority) class samples (treating them as noise) and predict the accuracy close to 100 percent of prevalent (majority) class. In the medical domain, the correct classification of samples of small class often has a greater value than the contrary case (cancerous patients classified as non cancerous).

Therefore, not only between-class imbalance generates an imbalance problem, but also data complexity, such as lack of data, class overlapping small disjuncts, noisy data and dataset shift also influence the classification accuracy.

1) Imbalance due to rare instances: The sample size plays an important role in determining the "effectiveness" of a classification model. A data set in which minority class samples are very limited is known as imbalance due to rare instances. The observation in Ref. [8] indicate that when the number of examples of the training set increases, the error rate caused by the imbalanced class distribution decreases. This problem is also related to the "lack of density" or "lack of information". When sample size is very small it is difficult for algorithms to distinguish rare examples from the prevalent class samples.

2) Class seperatibility or Overlapping: When a sample of one class overlaps on another class it is known as class overlapping. It is difficult to discriminate such kind of overlapping classes and therefore much harder rules are induced to discriminate such samples. When the samples are highly overlapped, it can significantly decrease the number of minority class examples correctly classified. The observation in ref. [8] state that "linearly separable" problem do not sensitive to any amount of imbalance

3) Small Disjunct: The existence of small disjuncts in data sets occurs when the concepts are represented within small clusters, where minority class is formed of sub concepts . The existence of sub concepts also increases the complexity of the problem because it becomes hard to know whether these samples represent actual sub concepts or are noise samples

4) Noisy Data: The presence of noise has a greater impact on the rare classes than on usual cases, since the rare class has fewer instances to begin with, it will take fewer "noisy" examples to impact the learned sub concept. In order to avoid the erroneous generation of discrimination functions for these "noise-areas" examples, some over fitting techniques must be employed, such as pruning. However, the shortcoming of this methodology is that some correct rare classes will be ignored and in this manner, the bias of the model should be set in order to be able to provide a good global behavior for both classes of the problem.

5) Dataset Shift: The problem of dataset shift is defined as the case where the training and test data follow different distributions. This problem often appears due to sample selection bias. The dataset shift problem is important when dealing with highly imbalanced domain, in which the minority class is sensitive to singular classification errors, due to the low number of examples it presents [9].

B. Reported Research solution

Due to the importance of the imbalanced dataset problem, a large number of solutions are reported in literature. These solutions are categorized into data level, algorithm level, cost-sensitive learning and ensemble learning, depending on how they deal with the problem.

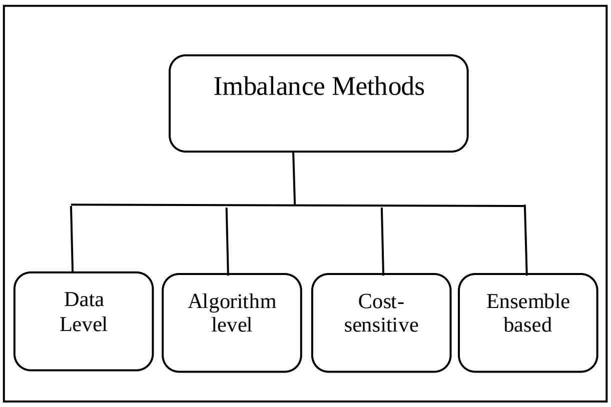

Figure 3. Class imbalance methods.

a) At the data level approaches (also called internal), the objective is to re-balance the class distribution by re-sampling the data space [10],[11],[12] [13]. The solution at the data level consists of the modification of an imbalanced data set by some mechanism in order to provide a balanced distribution. The main advantage of data level approach is that they are more versatile and their use is independent of underlying classifier. This category can be classified into three groups:

1. Under sampling methods create a subset of the original dataset by eliminating instances of the majority prevalent class.
2. Oversampling methods create a superset of the original data set by replicating some of the instances or creating new ones from the existing ones.
3. Hybrid methods that combine both oversampling and under sampling methods. This method increases the size of minority class while simultaneously decreasing the majority class.

b) At the algorithm level approaches (also called external), the solution tries to adapt existing classifier learning algorithms to bias the learning toward the minority class. This method creates new algorithms or modifies existing ones to take the class imbalance problem into consideration. [14],[15].

c) Cost-sensitive learning incorporates both data level and algorithm level approaches assuming higher misclassification costs with examples of the rare class with respect to the prevalent class, therefore, seek to minimize the high cost errors [16, 17] .

Cost-sensitive learning biases the classifier toward the rare class and therefore the rare class gain importance. The main feature of this method is that it tries to minimize the total cost of misclassification. In cost-sensitive methods it is more interesting to recognize the positive instances rather than the negative ones. For example, in medical domain the cost of misclassifying a non cancerous patient is limited to additional medical tests, while the cost of misdiagnosis will be fatal as potentially cancerous patients will be considered healthy. Therefore, the cost associated with a positive instance must be greater than the cost of misclassifying a negative one, i.e. C(+,-)>C(-,+).

The major drawback of this approach is that the costs are precisely unknown, and we usually tend to use approximations or ratios of proportionality

d) Ensemble based classifiers are designed to improve the accuracy of a single classifier by training several classifiers and combining them to output a new classifier that outperforms every one of them. Therefore, ensemble based methods are based on the combination between ensemble learning algorithms and one of the technique, such as data level and cost-sensitive ones. The main motivation for the combining classifiers in redundant ensemble is to improve their generalization ability: each classifier is known to make errors with the assumption that they have been learned on different data sets or they have different behavior over different part of input space; the examples that are misclassified by different classifier, however, are not necessarily the same [17].

IV. Related Work

In recent years, various techniques have been developed to deal with the class imbalance problem. In this section, we will discuss existing research on the classification of Imbalance datasets.

A. Data level approaches:

1) Random Oversampling (ROS)[18]: This is non-heuristic method that aims to balance class distribution through randomly replicating minority class examples. The disadvantage of this method is that it can increase the likelihood of occurring over-fitting/over-generalization, since it makes exact copies of existing samples.
2) Random Under-sampling (RUS)[18]: It is a non-heuristic method that aims to balance class distribution through randomly elimination of majority class instances. The major disadvantage of this method is that it can discard potentially useful data that could be important for learning process.
3) SMOTE (Synthetic Minority Over-sampling Technique [19]): This is an oversampling method; its main idea is to create new minority class samples by interpolating several minority class instances that lie together for oversampling the training set. In SMOTE, the minority class is over-sampled by taking each minority class examples and introducing synthetic examples along the line segments joining any/all of k nearest neighbors of a minority class. Depending upon the oversampling ratio, neighbors from the k nearest neighbors are randomly chosen. However, in the SMOTE algorithm, the problem of over generalization is avoided.

B. Ensemble learning:

The basic idea of ensemble learning is to try to improve the performance of single classifiers by inducing multiple classifiers and then aggregate their predictions to obtain a new classifier that outperforms every one of them. This technique follows the human natural behavior that tends to seek several opinions before making any important decisions. The keys for good performance of ensembles is the “diversity”, that is achieved by the combination between ensemble learning algorithms and one of the techniques, sampling or cost-sensitive learning solutions. The most widely used ensemble learning algorithms are Bagging and AdaBoost, which is most successful in variance reduction.

1) SMOTEBagging [19]: In Bagging, each base classifier is obtained from a random sample of training data. This method combines bagging with SMOTE and oversampling in each iteration, so that the data set is completely balanced. This procedure can be developed in two ways: (i) bootstrapped replica of the majority class instances; and (ii) using SMOTE and random oversampling depending on re-sampling rate.
2) SMOTEBoost [19]: In boosting, the whole training set is used to train each classifier. SMOTEBoost

introduce synthetic instances using SMOTE algorithm. Since new samples are created, the weights of the new samples are proportional to the total number of samples in the new data set. The weights of the samples from the original dataset are normalized in such a way that they form a distribution with the new samples.

3) RUSBoost [20]: RUSBoost removes instances from the majority class in each iteration using the random under sampling method. The weights of the remaining instances in the new datasets are normalized to form a uniform distribution.

V. Proposed Approach

In this section, we introduce the proposed algorithm for classification of imbalanced dataset.

*A. **5.1 Algorithm:** K-means clustering*

Input: Dataset consists of data and classes. Dataset “data” is divided into two classes. Two classes are ‘Negative class’as‘Major classes and ‘Positive class’as‘ Minor classes.

Output: Generate a balance dataset which would produce more effective results

Procedure:

```
1.  Read data belongs to negative class
        Ndata←data(class='negative');
2.  Divide Ndata into NC number of clusters
        IDX, Dist←CLUSTERING(Ndata,NC);
3.  For i=1 to length (IDX)
        distCentroid(i,IDX(i))←D(i,IDX(i));
    End for
4.  For i=1 to NC
        MaxDist(i)←max(distCentroid(:,i));
    End for
5.  For i=1 to length (IDX)
        distPercentile(i)←distCentroid(i,IDX(i))/
        MaxDist(IDX(i))*100;
    End for
6.  pickData←Ndata((distPercentile>=70),:);
7.  For i=1 to length(pickData)
        class1 (i)←'negative';
    End for
8.  Read data belongs to positive class
        Pdata←data(class='positive'),:);
9.  For i=1 to length(Pdata)
        class2(i) ← 'positive';
    End for
10. Merge two array length-wise
        datatrain=[pickData;Pdata];
11. Training the classifier
12. classtrain=[class1;class2];
13. Cls=j48(datatrain,classtrain);
14. Testing the classifier
15. Result=J48(Cls,data);
```

5.2 Algorithm:

Particle Swarm Optimization (PSO) Based

Input:

Dataset consists of data and classes. Dataset “data” is divided into two classes. Two classes are ‘Negative class’as‘Major classes and ‘Positive class’as‘Minor classes.

Output:

Generate a balance dataset which would produce more effective results.

Procedure:

```
PSO_Cluster(X, k, C)
        // X is data for clustering, k is number of cluster.
        // Let C be array of centroid randomly selected set,
// number of particle is pa.
        // Initialize swarmVel (k,X[0].length, pa) with random
// values between {0,1}
        // swarmPos = C is initial position of all particles,
// swarmBest is Best position initially null
        // swarmFitness(1:particles)=Inf is fitness value of
// each particle.

While (condition !=false)
        Distances←zeros(dataset_size(1),k,pa);
        for particle=1 to pa
        for centroid=1to k
        distance←zeros(dataset_size(1),1);

        fordataVector=1to X.length
        distance(data_vector,1)←EculedianDistance(swarm_
pos(centroid,:,particle)-X(data_vector,:));
        //Calculate Distance
        End for
        distances(:,centroid,particle)□distance;
        End for
        End for

        for particle=1:pa
        [value, index] ←min(distances(:,:,particle),[],2);
        partcalIndex(:,particle) ←index;
        End for

        FITNESS:=swarmFitness
        averageFitness = zeros(pa,1);

        for particle=1:pa
        for centroid = 1 : k
        if any(partcalIndex (:,particle) == centroid)
        localFitness=mean(distances(c(:,particle)==centroid,c
entroid,particle));
        averageFitness(particle,1) = averageFitness(particle,1)
+ local_fitness;
        End If
        End for
        End for

average_fitness(particle,1)= average_fitness(particle,1) / k;
```

```
if (average_fitness(particle,1) <swarm_fitness(particle))
        swarm_fitness(particle) = average_fitness(particle,1);
        swarm_best(:,:,particle) = swarm_pos(:,:,particle);
End if

[global_fitness, index] = min(swarm_fitness);
swarm_overall_pose = swarm_pos(:,:,index);
        for particle=1:pa
        inertia = w * swarm_vel(:,:,particle);
        cognitive = c1 * r1 * (swarm_best(:,:,particle)-swarm_pos(:,:,particle));
        social = c2 * r2 * (swarm_overall_pose-swarm_pos(:,:,particle));
        vel = inertia+cognitive+social;
swarm_pos(:,:,particle) = swarm_pos(:,:,particle) + vel ;
swarm_vel(:,:,particle) = vel;
End while
```

VI. Experimental Result

The performance of the proposed algorithm is evaluated in this section. The experiments were performed on a 2.20 GHz 32-Bits Intel i3 processor with 4 GB RAM. The operating system is Microsoft Windows 7. The algorithms are implemented in MATLAB 2014 with a WEKA package. The experiment was performed on the 17 most imbalanced binary data sets from the KEEL data-set repository and this dataset is downloaded from KEEL repository. The properties of the selected data-sets are shown in table 1. It shows, for each dataset, the number of examples (#Ex.), the number of attributes (#Atts) and the IR (Imbalanced Ratio).

The estimates of the accuracy rate (AUC) were obtained by using a 5-fold cross-validation, i.e., we split the dataset into 5 folds, each one containing 5% of the samples of the dataset. For each fold, the algorithms are trained with the samples contained in the remaining folds and then tested with the current fold. This process was carried out three times with different seeds. The data partitions used in this paper can be found in KEEL-dataset repository, so that any interested researcher can reproduce the experimental study.

TABLE I. Summary of Imbalance Data

Dataset	#Ex.	#Atts.	IR
Ecoli0137vs26	281	7	39.15
Ecoli0146vs5	280	6	13.00
Ecoli0147vs2356	336	7	10.59
Ecoli0147vs56	332	6	12.28
Ecoli01vs5	240	6	11.00
Ecoli0347vs56	257	7	09.28
Ecoli067vs5	220	6	10.00
Ecoli4	336	7	13.84
Glass016vs2	192	9	10.29
Glass016vs5	184	9	19.44
Glass4	214	9	15.47
Led7Digit02456789vs1	443	7	10.97
Pageblocks13vs4			
Shuttlec0vs4	1829	9	13.87
Shuttlec2vs4	129	9	20.50
Vowel0	988	13	10.10
Yeast2vs8	482	8	23.10

The Table II shows the accuracy achieved of both the methods used in the proposed work.

TABLE II. Accuracy of Clustering

Dataset	Existing Work	Work-1(kmeans)	Work-2 (PSO-Cluster)
Ecoli0137vs26	0.836	0.91303	0.91303
Ecoli0146vs5	0.9295	0.892308	0.92471
Ecoli0147vs2356	0.8943	0.900974	0.969551
Ecoli0147vs56	0.8924	0.953463	0.973407
Ecoli01vs5	0.9235	0.910739	0.969159
Ecoli0347vs56	0.8928	0.886307	0.933534
Ecoli067vs5	0.89	0.964796	0.964796
Ecoli4	0.9303	0.97053	0.97053
Glass016vs2	0.7488	0.661786	0.76978
Glass016vs5	0.9886	1	0.846154
Glass4	0.9192	0.813225	0.933333
Led7Digit02456789vs1	0.888	0.900487	0.900487
Pageblocks13vs4	0.9906	0.9375	0.982759
Shuttlec0vs4	1	1	1
Shuttlec2vs4	1	1	1
Vowel0	0.9887	0.99327	0.994505
Yeast2vs8	0.8019	0.594203	0.948759

VII. Conclusion

This paper explores the effects of imbalanced datasets. This article has addressed the class imbalance problem and its solution. This article suggests the solution of the data imbalance by the means of coupling of Clustering. There are two kind clustering techniques are implemented. First one is k-means clustering and in another clustering is done by Particle Swarm Optimization Techniques. Furthermore, it proposed a new ensemble combination technique that takes into consideration the accuracies for each class. It has been applied this method on synthetic as well as real datasets. The experimental results suggest that the coupling

REFERENCES

[1] A. Jain, M. Murty and P. Flynn, "Data Clustering: A Review", ACM Computing Surveys, Vol.31, No. 3, Sep 1999, pp. 264–323.

[2] G Ball, D Hall, "A Clustering Technique for Summarizing Multivariate Data", Behavioral Science, Vol. 12, pp 153–155, 1967

[3] Romero, C., & Ventura, S. Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 40(6), 601- 618 (2010).

[4] Han Jiawei, M. Kamber, "Data mining: Concepts and techniques" in , Beijing: China Machine Press, 2006.

[5] MR Rao, "Cluster Analysis and Mathematical Programming", Journal of the American Statistical Association, Vol. 22, pp 622-626, 1971.

[6] E Forgy, "Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of Classification", Biometrics, Vol. 21, pp 768–769, 1965.

[7] Tom Fawcett, Ira Haimowitz, Foster Provost, and Salvatore Stolfo "AI Approaches to Fraud Detection and Risk Management" AI Magazine Volume 19 Number 2 pp 107-108 (1998)

[8] Yanminsun & Wong, Andrew & Kamel, Mohamed S.. Classification of imbalanced data: a review. International Journal of Pattern Recognition and Artificial Intelligence. Vol.23, No. 4 pp 687-719 (2011).

[9] Kubat, M. and Matwin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One Sided Selection. In Proceedings of the Fourteenth Intemational Conference on Machine Learning, pages 179-186, Nashville, Tennesse. Morgan Kaufmann.

[10] N. V. Chawla, "Data mining for imbalanced datasets: An overview," in Data Mining and Knowledge Discovery Handbook, 2010, pp. 875–886.

[11] G. M. Weiss and F. Provost, "Learning when training data are costly: The effect of class distribution on tree induction," J. Artif. Intell. Res., vol. 19, pp. 315–354, 2003.

[12] N. Japkowicz and S. Stephen, "The class imbalance problem:Asystematic study," Intell. Data Anal., vol. 6, pp. 429–449, 2002.

[13] D. A. Cieslak and N. V. Chawla, "Start globally, optimize locally, predict globally: Improving performance on imbalanced data," in Proc. 8th IEEE Int. Conf. Data Mining, 2009, pp. 143–152.

[14] D. A. Cieslak and N. V. Chawla, "Start globally, optimize locally, predict globally: Improving performance on imbalanced data," in Proc. 8th IEEE Int. Conf. Data Mining, 2009, pp. 143–152.

[15] T. K. Ho, "Multiple classifier combination: Lessons and next steps," in Hybrid Methods in Pattern Recognition, Kandel and Bunke, Eds. Singapore: World Scientific, 2002, pp. 171–198.

[16] N. Ueda and R. Nakano, "Generalization error of ensemble estimators," in Proc. IEEE Int. Conf. Neural Netw., 1996, vol. 1, pp. 90–95.

[17] X. Hu, "Using rough sets theory and database operations to construct agood ensemble of classifiers for data mining applications," in Proc. IEEEInt. Conf. Data Mining, 2001, pp. 233–240.

[18] L. I. Kuncheva, "Diversity in multiple classifier systems," Inf. Fusion, vol. 6, no. 1, pp. 3–4, 2005 (diversity in multiple classifier systems).

[19] 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, "SIGKDD Expl. Newslett., vol. 6, pp. 20–29, 2004.

[20] N. V. Chawla, K.W. Bowyer, L. O. Hall, and W. P. Kegel meyer, "SMOTE: synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16,pp. 321–357, 2002.

[21] C. Seiffert, T. Khosh goftaar, J. Van Hulse, and A. Napolitano, "Rusboost: A hybrid approach to alleviating class imbalance," IEEE Trans. Syst. ,Man, Cybern. A, Syst., Humans, vol. 40, no. 1, pp. 185–197, Jan. 2010.

Dr. Mohd Ashraf pursed Bachelor of Technology in Computer Engineering from Gobind Bhallabh Pant University of Agriculture & Technology, Pantnagar (UK) in year 2004 and Master of Technology in Computer Science & Engineering , Aligarh Muslim University, Aligarh India in year 2009. He completed his Ph.D in the field of Computer Science & Engineering from Gautam Budhha University in year 2014. and currently working as Associate Professor in Department of Computer Science & Engineering, Maulana Azad National Urdu University , Hyderabad since 2015. He has published more than 40 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE and it's also available online. His main research work focuses on Optimization Algorithms, Network Reliability, Soft Computing, Fuzzy logic. He has 14 years of teaching experience and 5 years of Research Experience.

R. Mishra is an assistant professor at the School of ICT, Gautam Buddha University, Greater Noida (U.P.) India. He completed his BE (electronics engineering, 2000), MTech (Reliability Engineering, 2004), and PhD from the Reliability Engineering Centre, IIT-Kharagpur, in the year 2009. His current research interests are network reliability prediction, analysis, and design. He has published more than 20 research papers in various international journals and conferences such as International Journal of Performability Engineering (IJPE), IEEE Trans on Reliability, QTQM, and RESS. He is an editorial board member of IJEI and also a reviewer for IJSA.