International Journal on Applications in Information and Communication Engineering
Volume 5 : Issue 2 : December 2019 , pp 43 –49 www.aetsjournal.com ISSN (Online) : 2394 - 6237

CLUSTER CONCENTRIC CIRCLES BASED UNDERSAMPLING TO HANDLE UNBALANCED DATA

Mehala G, Bharathi S, Gajalakshmi S
Department of Information Technology
Rathinam Technical Campus, Coimbatore, Tamilnadu, India

Abstract— The most recent topic to be covered in the field of data mining is dealing with datasets that have an uneven distribution of classes. Conventional classification algorithms always aim to achieve the highest possible overall accuracy, but they do so without taking into account how the data are distributed within their own classes. This research offers a solution to the problem of an unevenly distributed dataset by developing an innovative cluster-based under-sampling method known as Cluster Concentric Circle based Under Sampling (C3BUS). C3BUS chooses the selective data as the training data in order to increase the effectiveness of the classifier and reduce the impact of an unbalanced distribution. Experimentation was performed on a synthetic dataset, as well as the Abalone, Bioassay, Glass, and Ecoli datasets, and the results of those experiments are presented here in order to demonstrate that the method that was suggested is effective. These findings take into account a number of different evaluation criteria, including accuracy, precision, sensitivity, specificity, F-measure, and time.

Index Terms— Classification, unbalanced data, sampling, cluster-based under-sampling, and a balanced dataset

I. INTRODUCTION

The availability of raw data has vastly increased the number of topics that can be researched within the realm of knowledge discovery [6]. One well-known approach to data mining is known as classification. Traditional classifiers make the assumption that the data used to train the classifier is evenly distributed across all of the classes, despite the fact that many datasets derived from the real world are imbalanced, which has a negative impact on the performance of the classifier. Data sets are considered to be imbalanced when they display an unequal distribution across the classes [10]. The requirement for a balanced dataset becomes apparent when classifiers have a tendency to favour the majority class over the minority class, which is more important to take into account. Traditional classification methods, which have a tendency to favour the frequently occurring cases (majority class) notwithstanding the high cost of incorrectly identifying the rarely happening examples (minority class), may only produce a suboptimal classification model for an unbalanced dataset [3]. Traditional classification methods have a tendency to favour the frequently occurring cases (majority class) notwithstanding the high cost of incorrectly identifying the rarely happening examples (minority class). The researchers have a predisposition to favour this problem due to the fact that it is so prevalent in a wide variety of applications in the real world. It is necessary to strike a balance between the classes because the conventional methods of classification all give an advantage to the group that constitutes the majority, which will result in improved performance.

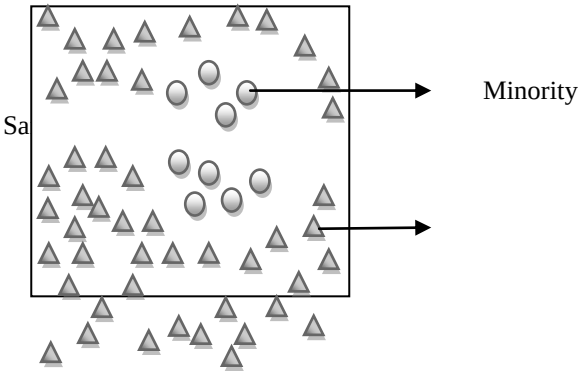

Fig. 1a. Imbalanced Dataset

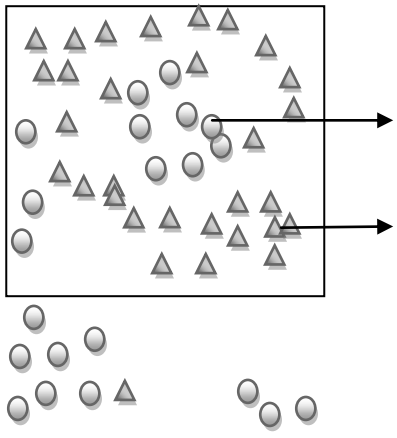

Fig. 1b. Balanced Dataset

There are three different approaches to handling datasets that are not evenly distributed. First, data sampling 2. Managing algorithms, and 3. Education that takes cost into account [19]. In the first method, the training cases are re-sampled in order to produce a balanced distribution. The second solution addresses the issue by either developing a brand-new algorithm or modifying an existing one. Increasing the cost of cases that were incorrectly identified as positive is part of the third strategy, which combines data-level, algorithmic-level, or hybrid-level approaches. The artificial dataset that was produced by combining these three methods is different from the distribution that was originally used. The aforementioned dataset can therefore be processed using conventional techniques; however, given that the test points originate from the original distribution, there is a possibility of a disagreement between the test points and the original points [20]. [Citation needed] Data sampling either adds samples to the rare class (also known as oversampling) or subtracts samples from the frequently occurring class (also known as undersampling), both of which have their own set of benefits and drawbacks [3,19]. Under-sampling occurs when some of the samples in a dataset are removed so that it has a more even distribution. This can result in the loss of important information. In addition, because the size of the dataset is reduced, the amount of time necessary to train the samples is reduced as well. RUS is the simplest form of under-sampling, and it involves removing samples at random from the majority class in order to maintain the distribution's equilibrium [11, 22]. Oversampling eliminates the problem caused by undersampling, but it obviously extends the amount of time necessary to train the model because it either duplicates the data or adds new ones in order to achieve a more even distribution. An efficient classifier can be developed from an unbalanced dataset by employing a selection strategy that either under- or oversamples the majority class and the minority class. An unbalanced set of data may result in a categorization model that completely disregards the minority class. For a Classifier, a two-by-two confusion matrix is used as the foundation for determining criteria such as the True positive rate, the False positive rate, Sensitivity, and Specificity [17].

This research endeavour is organised as follows in order to accomplish its primary objective, which is to create an efficient classifier model despite the presence of an imbalanced dataset. The Literature Survey for Part II is discussed in this section. In Section III, the C3BUS technique that has been suggested can be seen. This investigation draws to a close with the results of the experiments discussed in Sections IV and V.

2. Extra Labor Required

Recent developments in science and technology have made it possible to collect unbalanced datasets, and this has piqued the interest of a significant number of academics. In the field of data mining, one of the new challenges that needs to be addressed is the problem of imbalanced datasets [15]. It has been suggested to use a method of random undersampling that is determined by distance in order to achieve distributional equilibrium among the classes. In their conclusion, the authors state that the performance of classification systems is improved when balanced datasets are used. When it comes to clinical datasets, under-sampling performs significantly better than over-sampling and has higher recall rates [11]. M. Mostafizur Rahman and his colleagues conducted research on a cluster-based under-sampling strategy in order to address the problem of an uneven distribution of classes. His findings show that the suggested strategy performs noticeably better than the existing cluster-based undersampling methods [9]. [Since his findings demonstrated this, the suggested strategy] Cluster-based under-sampling methodologies were presented by Show-Jane Yen and Yue-Shi Lee. The goals of these methodologies were to reduce the impact of imbalanced datasets and improve the accuracy of predictions regarding minority class. This author demonstrated the superiority of the suggested method when compared to the performance of the existing methods by using synthetic datasets in addition to two more datasets that were obtained from the repository of datasets [16]. The researchers Parinaz, Herna, and their coworkers looked into a single classifier that chose the samples by employing a centroid-based cluster under-sampling strategy. The author believes that cluster centroids do not provide any useful information. In his second experiment, he looked into the ClusFirstClass undersampling ensemble technique, which is considered to be superior to the other innovative available choices [13]. Rushi Longadge and colleagues came up with the idea for the multi cluster-based majority under-sampling strategy. Cluster-based random undersampling, as opposed to undersampling, is a method that has the potential to successfully prevent the vital information belonging to the majority class from being lost [8]. Chris Seiffert and his colleagues have developed a novel sampling and boosting technique combination that they call RUSBoost. The authors demonstrated that RUSBoost is a compelling alternative to conventional algorithms by evaluating the performances of RUSBoost, SMOTEBoost, and each of their individual components (random undersampling, SMOTE, and AdaBoost) [3.]

This allowed the authors to demonstrate that RUSBoost is a compelling alternative to conventional algorithms. Kai-Biao Lin et al. [7] presented a novel method known as FCM-SVM, which not only improved minority recall but also ensured excellent classification accuracy for the minority class.

There were two different decision tree ensembles that were suggested by Yubin Park and Joydeep Ghosh. The first ensemble was successful in producing a wide variety of decision trees by making use of novel splitting criteria that were determined by the alpha divergence. The second ensemble used the same alpha trees as the base classifiers, but it used a lifting aware stopping criterion to halt the growth of the tree. This resulted in the production of a set of understandable rules that increased the lift values [20]. The Synthetic Minority Over-sampling Technique (SMOTE) approach, which was put forward by Chawla et al., oversamples minority samples by constructing artificial examples. This method is in contrast to the practise of duplication. The minority class is oversampled by including synthetic samples along the line segments connecting any and all of the k adjacent minority class neighbours. The neighbours are chosen at random from the k closest neighbours, and the amount of oversampling that is required is taken into consideration [2].

3. A methodology known as C3BUS, which is an acronym for Cluster Concentric Circle Based Under-Sampling

The cluster-based undersampling method [9] is being used currently, but the method that is being presented in this work is different from that method. The primary emphasis of this work is on finding a way to select the cases in such a way as to prevent undersampling, which is the primary challenge presented by this sampling strategy, and the omission of any instances that may be potentially significant. An unbalanced dataset, on the other hand, does not have a uniform distribution of samples across all of the classes, in contrast to a balanced dataset. Using a sampling methodology, this method is modelled by constructing a balanced dataset out of an unbalanced one. This dataset was initially unbalanced. Two methods of sampling are referred to as undersampling and oversampling respectively. Random under sampling and focused under sampling are the two categories that fall under the umbrella term "under sampling." In this case, cluster-based undersampling is taken into account and improved upon in order to ensure that the selection procedure does not fail to take into account the required sample. The dataset DS is partitioned into two groups: those that represent the majority (DSM) and those that represent the minority (DS) (DSm). $DS = DS_M + DS_m$ (1)

The majority samples are grouped into different k clusters using K means algorithm. This clustering algorithm is computationally faster than other hierarchical clustering algorithms [23].

$$DS_M = \bigcup_{i=1}^{k} DSC_i \quad (2)$$

The number of samples to be chosen from i^{th} cluster is calculated using (3) The lower bound of i is set to 1 and upper bound to k.

$$nc_i = \frac{|DS_m|}{|DS_M|} \times |C_i| \quad (3)$$

For all the clusters calculate

$$d (Cc, S_i) \text{ where } i = 1.....K \quad (4)$$

Where Cc is the cluster center and S_i represents the samples in the i^{th} cluster. Each cluster is then divided into nc_i concentric circles with the distance of n (calculated using 5). np and fp is considered as the nearest and farthest sample to the cluster center respectively.

$$n = d (np , fp)/ nc_i \quad (5)$$

First sample is chosen using (6) and the further samples are chosen from each concentric circle using (7). The cluster is divided into concentric circles in such a way that one sample is chosen from each circle. The chosen samples are then combined with minority class to form a balanced dataset (BDS).

$$SC_j = \text{np (with j as 1)} \quad (6)$$

Where np is the nearest sample in the cluster from the centroid. SC_j is the first chosen sample in the first concentric circle of a cluster which is the nearest sample from the cluster center. Next sample is selected in such a way that the sample is the farthest sample to the previously selected one in the next concentric circle. Further samples are selected in the same manner until the count reaches nc_i.

$$SC_{j+1} = \text{Max}\{ d (SC_j , SCC_j) \} \quad [\text{where j= 1 to } nc_i] \quad (7)$$

SCC_j represents the samples in the next concentric circle of a cluster.

The same process is repeated for each cluster and then the chosen samples are combined with the minority samples to for a balanced dataset as in (8). Finally the balanced dataset is trained with neural network classifier. Fig.2 pictorially represents flow of the proposed methodology.

$$BDS = DS_m \cup (SC_j) \quad (8)$$

4. Laboratory Analysis

In this section, the suggested C3BUS strategy is evaluated using both synthetic and repository datasets in order to determine how successful it was. Using a neural network classifier, a comparison is made between the performance of the C3BUS and the currently used cluster-based under-sampling techniques. A large number of researchers in the academic community (including R. Barendela et al., 2003; F.J. T. Fawcett Provost (1997), F. J. Provost et al. (1998), and N.V. According to Chawla and colleagues, overall accuracy is not the only suitable evaluation parameter for a classifier when the dataset it is being applied to is unbalanced. because the breakdown of the samples into their respective classes is not taken into consideration. In this investigation, the C3BUS, the existing Cluster Based Under-sampling approach (APP1), and the initial imbalanced dataset are each evaluated based on their performance according to one of five metrics. Accuracy, Precision, or Positive Predicted Value (V. Garcia et al.,), Sensitivity or Recall, or True Positive Rate, Specificity, and F-measure or F-Score are the names of these metrics (ODS). In addition to these, the passage of time is also taken into consideration. Using a Neural Network classifier in a MATLAB environment, the proposed C3BUS is tested against a synthetic dataset in addition to the Abalone, Bioassay, Glass, and Ecoli datasets. The results of this test are then analysed. It can be inferred that the imbalance ratio has some bearing on the classifier's overall effectiveness. Altering the value of the K parameter in the K-means algorithm will result in different outcomes. The proportion of data points that come from underrepresented classes has a sizeable impact on the overall balance of the dataset. When a minority group only has a very few samples to work with, it is difficult to ensure that the dataset is balanced. When compared to the execution time of the previous approach, the time required by the proposed method to classify the dataset is noticeably less.

A. Results Obtained from Synthetic Datasets

The performance of C3BUS is evaluated by first creating a synthetic dataset [15] with an imbalance ratio of 10:1. This dataset contains a total of 11000 instances, 1000 positive samples, and 5 attributes. Positive samples make up 1000 of the total. It has been found that C3BUS performs better than both the approach that is currently being used and the dataset that was originally imbalanced. Fig. 1 and Table I. Figure 3 displays the results of analyses conducted on artificial datasets.

Results of NN Running on the Synthetic Dataset Shown in Table I

Method/Metrics	**APP1 (%)**	**ODS (%)**	**C3BUS (%)**
Accuracy	48	75	91
precision	57	76	90
sensitivity	71	73	90
specificity	28	26	9
F-measure	25	73	91

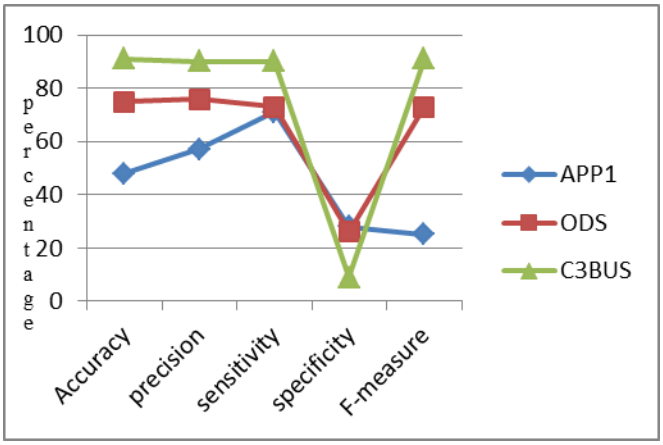

Fig 3. Comparison of measures on Synthetic dataset

B. Findings from the Abalone Dataset

The C3BUS performance is further evaluated using the Abalone dataset, which is retrieved from the KEEL repository. The ratio of imbalanced data in this dataset is 128 to 1.

Each of the 4174 cases that make up this dataset has nine properties. The findings from the Abalone dataset that was used in this experiment are consistent with the findings that the proposed strategy offers superior performance, which is the case. Table II and Figure 4 both illustrate the performance that was achieved.

Table II. Performance of NN on Abalone Dataset

Method/Metrics	APP1 (%)	ODS (%)	C3BUS (%)
Accuracy	45	67	87
precision	41	62	86
sensitivity	43	64	89
specificity	56	35	10
F-measure	54	63	88

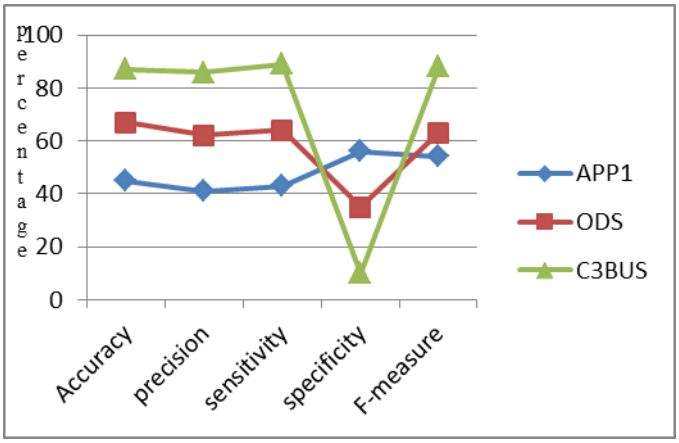

Fig 4. Comparison of measures on Abalone dataset

C. Results from the Bioassay Dataset

The effectiveness of C3BUS is analysed by making use of the Bioassay dataset, which consists of 3441 instances of the majority class and 60 instances of the minority class, along with 145 attributes. This particular set of data has an imbalance ratio of 57 to 1. The results of this experiment's dataset show that the proposed strategy performs more effectively than its predecessor. Table III and Figure 5 present the results that were obtained from the Bioassay dataset.

Table III Performance of NN on Bioassay Dataset

Method/Metrics	APP1 (%)	ODS (%)	C3BUS (%)
Accuracy	60	74	90
precision	80	72	90
sensitivity	50	73	91
specificity	50	26	8
F-measure	70	72	88

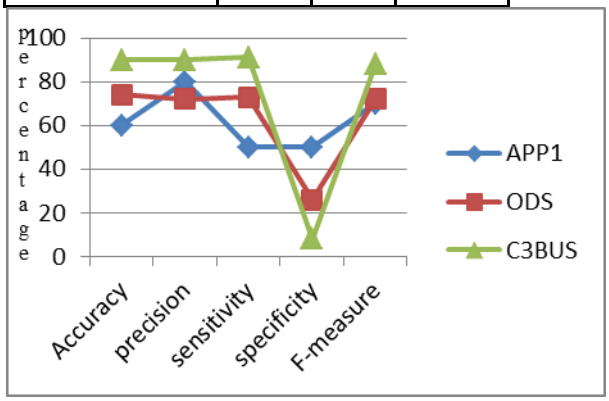

Fig 5. Comparison of measures on Bioassay dataset

D. Results on Glass Dataset

This data is collected from KEEL repository with 35% of minority positive samples and 65% of majority negative samples. It has an imbalance ratio of 2:1. Table IV and Fig 6 proves that C3BUS surpluss the other methods.

Table IV Performance of NN on Glass Dataset

Method/Metrics	APP1 (%)	ODS (%)	C3BUS (%)
Accuracy	93	58	95
precision	93	71	95
sensitivity	93	70	95
specificity	6	30	5
F-measure	76	41	94

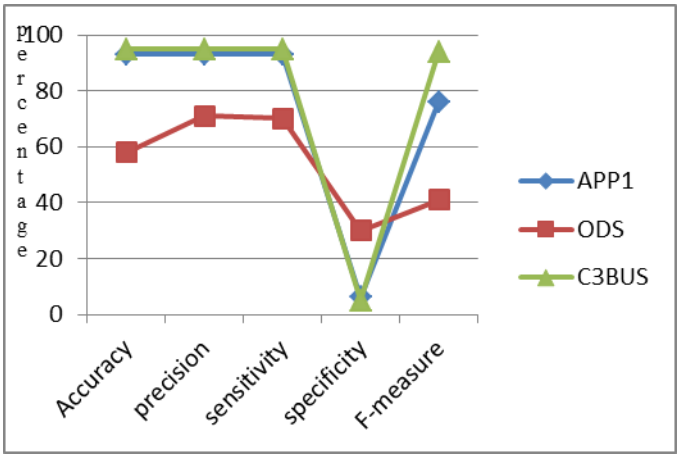

Fig 6. Comparison of measures on Glass dataset

E. Results on Ecoli Dataset

Ecoli dataset is obtained from KEEL repository with 7 features and an imbalance ratio of 1.86. It includes 34% of positive minority samples and 65% of negative majority samples. Table V and Fig 7 shows the results.

Table V Performance of NN on Ecoli Dataset

Method/Metrics	**APP1 (%)**	**ODS (%)**	**C3BUS (%)**
Accuracy	50	79	96
precision	52	82	96
sensitivity	50	83	96
specificity	50	16	3
F-measure	66	64	96

Table VI. Execution time for Existing approach and C3BUS

Method / Dataset	**APP1 (sec)**	**C3BUS (sec)**
Synthetic	75	3
Abalone	37	2
Bioassay	250	10
Glass	5.5	1
Ecoli	4	0.7

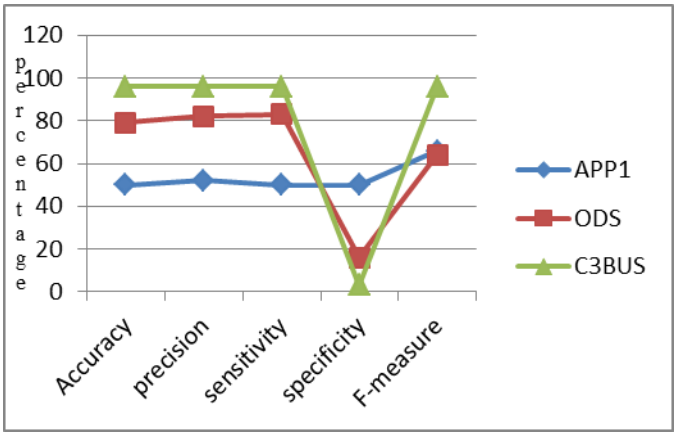

Fig 7. Comparison of measures on Ecoli dataset

The following table VI and Fig 8 shows the time taken to classify the dataset with the existing method and C3BUS.

5. CONCLUSION

When there is an imbalance in the distribution of the classes, it is more challenging for the classifier to correctly assign a category to the group that is underrepresented. In order to address the problem of class disparity, numerous potential solutions have been proposed. The Cluster Concentric Circle based under-sampling (C3BUS) technique was proposed in this paper as a means of bringing the unbalanced dataset into equilibrium. After that, the balanced dataset is used as an input for the neural network classifier, which sorts the samples into the appropriate categories. This work is compared to a different cluster-based under-sampling approach, and it is shown to perform better than the previous work in terms of accuracy, precision, sensitivity specificity, F-measure, and execution time for the datasets that were chosen. [Cluster-based under-sampling] It is possible that the suggested method could be improved even further by employing a hybrid sampling methodology to achieve dataset parity.

REFERENCES

[1]. [1] A correlation study on Hepatitis C virus load determined by real-time polymerase chain reaction with blood biomarkers in patients with renal illness was published in J Mol Biomark Diagn Vol 3 and Issue 2 in 2012. Bagyalakshmi R, Malathi J, Prathiba K, Samson Y, Ravichandran R, and Madhavan HN.

[2]. "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research 16, pp. 321-357, 2002. Chawla, N., Bowyer, K., Hall, L., and Kegelmeyer, W

[3]. "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance," IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, vol. 40, no. 1, January 2010. Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano.

[4]. F.J. Provost and T. Fawcett, "Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions," Proc, International Conference on Knowledge Discovery and Data Mining, pp. 43–48, 1997.

[5]. T. Fawcett, R. Kohavi, and F. J. Provost, "The Case Against Accuracy Estimation for Comparing Induction Algorithms," Proc. Int'l Conf. Machine Learning, pp. 445–453, 1998.

[6]. Haibo He, "Learning from Imbalanced Data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, September 2009.

[7]. "Imbalance Data Classification Algorithm based on SVM and Clustering Function," The 9th International Conference on Computer Science & Education (ICCSE 2014), Vancouver, Canada, August 22–24, pp. 544-548. Kai-Biao Lin, Wei Weng, Robert K. Lai, and Ping Lu.

[8]. "Multi-Cluster Based Approach for Skewed Data in Data Mining", IOSR Journal of Computer Engineering (IOSR-JCE), pp. 66-73, July-August 2013, by Mr. Rushi Longadge, Ms. Snehlata S. Dongre, and Dr. Latesh Malik.

[9]. M. Mostafizur Rahman and D. N. Davis, "Cluster based undersampling for unbalanced Cardivascular data," WCE 2013, July 3-5, 2013, Proceedings of the international congress on engineering, Vol. III.

[10]. Mikel Galar, A. Fernandez, E. Barrenechea, and H. Bustin, "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 and Issue 4, pp. 463–484, 2011.

[11]. "Balancing Class for Performance of Classification with a Clinical Dataset," Proceedings of the World Congress on Engineering 2014 Vol. I, July 2–4, 2014, London, U.K.; N. Poolsawad, C. Kambhampati, and J.G.F. Cleland.

[12]. "SMOTEBoost: Improving Prediction of the Minority Class in Boosting," Proc. Seventh European Conf. Principles and Practice of Knowledge Discovery in Databases, pp. 107-119, 2003. N.V. Chawla, A. Lazarevic, L.O. Hall, and K.W. Bowyer.

[13]. "Learning from Imbalanced Data Using Ensemble Methods and Cluster-based Undersampling," New Frontiers in Mining Complex Patterns, by Parinaz Sobhani, Herna Viktor, and Stan Matwin 2015 April, pages 69–83 of Lecture Notes in Computer Science Volume 8983.

[14]. Q. Yang and X. Wu, "10 difficult issues in data mining research," International Journal of Information Technology and Decision Making, volume 5, issue 4, pages 597–604, 2006.

[15]. Strategies for Learning in Class Balance Problems, R. Barendela, J. S. Sanchez, V. Garcia, and E. Rangel, Pattern Recognition 36, 849–851, 2003.

[16]. Yue-Shi Lee and Show-Jane Yen, "Cluster-based under-sampling techniques for imbalanced data distributions," Expert Systems with Applications, Vol. 36, Issue 3, Part I, 2009, pp. 5718–5727.

[17]. An introduction to ROC Analysis by Tom Fawcett, Elsevier Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

[18]. Sotoca and Mollineda, V.G.J.S.R. (2007). The problem of class imbalance in pattern learning and categorization. The document was downloaded from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.329.4200&rep=rep1&type=pdf.

[19]. Victoria Lopez, A. Fernandez, S. Garcia, V. Palade, and F. Herrera, "An insight into categorization with unbalanced data: Empirical results and current trends on exploiting data inherent characteristics", Elsevier Inc., Volume 250, pp. 113–141, Nov 2013.

[20]. Yubin Park and Joydeep Ghosh, "Ensembles of alphaTrees for Imbalanced Classification Problems," IEEE transactions on knowledge and data engineering, Vol. 26, No. 1, January 2014, pp. 131–143.

[21]. Cost-Sensitive Boosting for Classification of Imbalanced Data, Y. Sun, M.S. Kamel, A.K.C. Wong, and Y. Wang, Pattern Recognition, vol. 40, pp. 3358-3378, 2007.

[22]. "Cluster based majority under-sampling techniques for class imbalance learning," 2nd IEEE International Conference on Information and Financial Engineering (ICIFE), pp. 400–404, 2010. Z. Yan-ping, Z. Li-Na, and W. Yong-Cheng.

[23]. http://playwidtech.blogspot.in/2013/02/k-means-clustering-advantages-and.html.