Middle-East Journal of Scientific Research
24 (Special Issue on Innovations in Information, Embedded and Communication Systems): 314-319, 2016
ISSN 1990-9233; © IDOSI Publications, 2016
DOI: 10.5829/idosi.mejsr.2016.24.IIECS.23179

Cluster Concentric Circle Based Undersampling to Handle Imbalanced Data

[1]S. Srividhya and [2]R. Mallika

[1]Research and Development Centre, Bharathiar University, Coimbatore - 46, Tamilnadu, India
[2]Department of Computer Science, C.B.M College, Coimbatore - 42, Tamilnadu, India

Abstract: The most emerging problem in data mining is dealing the datasets with imbalanced class distribution. All the traditional classification algorithms aim to optimize the overall accuracy without taking into account the distribution of data in its classes. This paper proposes a solution to the imbalanced dataset by introducing a new cluster based under-sampling method called Cluster Concentric Circle based Under Sampling (C3BUS). C3BUS picks up the selective data as the training data to maximize the efficiency of the classifier and to minimize the influence of imbalanced distribution. Experimental result on a synthetic dataset, Abalone, Bioassay, Glass and Ecoli datasets are provided to be evidence for the effectiveness of the proposed method by considering Accuracy, Precision, Sensitivity, Specificity, F-measure and time as an evaluation measure.

Key words: Classification · Imbalanced data · Sampling · Cluster based Under-sampling · Balanced dataset

INTRODUCTION

Availability of raw data has created a colossal opportunity in the field of research in knowledge discovery [1]. One of the well known techniques in Data Mining is Classification. Traditional classifiers assume that the data used to train the classifier is balanced between the classes, but many real world datasets are imbalanced which degrades the performance of classifiers. Datasets which exhibit unequal distribution between the classes are called as Imbalanced datasets [2]. The importance of a balanced dataset is recognized only when the classifiers tend to bias towards the majority class and ignores the minority class which is more important to be considered [3]. With Imbalanced dataset, only a sub optimal classification model will be created by using traditional classification algorithms which tends to favour the frequently occurring examples (majority class) even though the cost of misclassifying the rarely occurring examples (minority class) is very high [4]. Researchers are biased towards this issue due to its presence in many real world applications. Since all the traditional classification algorithms favour the majority class there is a need to balance the classes to improve the performance.

The imbalanced datasets can be handled in three ways 1. Data Sampling 2. Algorithmic handling and 3. Cost Sensitive learning [5]. The first method re-samples the training instances to produce a balanced distribution [6]. The second one either develops a new algorithm or modifies an existing one to handle the issue. The third method incorporates data level, algorithmic level or hybrid level, by assigning higher cost to misclassified positive instances. These three methods create an artificial dataset which is different from the original distribution. So traditional algorithms can be applied to the above dataset but however the test points are from the original distribution which might cause discrepancy between test points and original points [7]. Data sampling handles the class distribution either by adding samples to the rare class (over sampling) or by removing samples from the frequent occurring class (under-sampling) with their advantages and drawbacks. Under-sampling might lose required information when samples are removed to balance the dataset. At the same time, it consumes less training time to train the samples since the size of the dataset is reduced. The simplest form of under-sampling is RUS which randomly removes samples from majority class to balance the distribution [8]. On the other hand, Oversampling retains all the samples and overcomes the drawback of under-sampling but obviously takes a longer time to train the model since it duplicates samples or creates new ones to balance the distribution. With the presence of imbalanced dataset, an efficient classifier can be built by the selection strategy of the majority class and

Corresponding Author: S. Srividhya, Research and Development Centre, Bharathiar University, Coimbatore - 46, Tamilnadu, India.

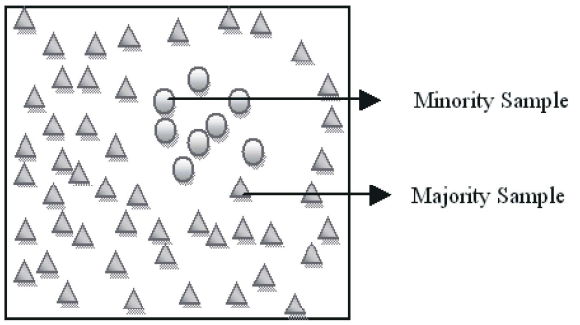

Fig. 1a: Imbalanced Dataset

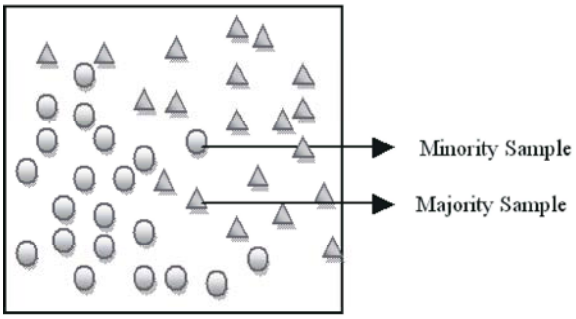

Fig. 1b: Balanced Dataset

minority class to under-sample or oversample [9]. A classification Model built with imbalanced data might completely ignore the minority class. For a Classifier, a two-by-two confusion matrix form the basis for the metrics like True positive rate, False positive rate, Sensitivity, Specificity [10].

With an aim to build an efficient classifier model in the presence of imbalanced dataset, this paper is organized in the following manner. Section 2 presents the Literature Survey. Section 3 shows the proposed C3BUS method. Finally this paper ends up with experimental results and conclusion in Section 4 and 5.

Related Work: Recent developments in Science and Technology paved the way for imbalanced datasets which grabbed the attention of many researchers. The imbalanced dataset is one of the emerging problems in Data mining which need to be considered [11]. Distance based random under-sampling method is proposed to balance the class distribution. The authors conclude saying that the performance of classification algorithms are better with the balanced datasets. Under-Sampling provides better recall rates than over-sampling and it performs better for Clinical datasets [12]. M.Mostafizur Rahman *et al*., explored cluster based under-sampling method to alleviate the imbalanced class distribution. His results prove that the proposed method shows significantly better performance than the existing cluster based under sampling methods [13]. Show-Jane Yen, Yue-Shi Lee proposed cluster based under-sampling approaches to decrease the influence of imbalanced datasets which in turn can increase the prediction of minority class. This author proved that the proposed method excels, comparing the performance of the existing methods by using synthetic datasets and two more datsets from the dataset repository [14]. Parinaz, Herna *et al*., explored a single classifier which used centroid based cluster under-sampling method to choose the samples. The author reports cluster centroids are not informative. In his second experiment, he explored undersampling ensemble algorithm based on clustering called ClusFirstClass which outperforms the other state of art solutions [15]. Rushi Longadge *et al*., proposed multi cluster-based majority under-sampling approach. Comparing with the under-sampling, cluster-based random under-sampling can effectively avoid the important information loss of majority class [16]. Chris Seiffert *et al*., presents RUSBoost algorithm which is new hybrid of sampling and boosting technique. The authors evaluates the performances of RUSBoost, SMOTEBoost and their individual components (random undersampling, SMOTE and AdaBoost) and proved that RUSBoost is an attractive alternate for traditional algorithms. Kai-Biao Lin *et al*., [17] proposed a new algorithm FCM-SVM which ensured high classification accuracy of minority class and also increases the recall of minority. Barendela *et al*. removes the noisy instances from the majority class with Wilson's algorithm [18].

Yubin Park and Joydeep Ghosh proposed two decision tree ensembles. First ensemble used novel splitting criteria based on alpha divergence which created diverse decision trees in the ensemble. The second ensemble used the same alpha trees as base classifiers but used a lift aware stopping criterion to stop the tree growth this provides set of interpretable rules which improved the lift values [7]. Chawla *et al*., proposed Synthetic Minority Over-sampling Technique (SMOTE) approach in which the minority samples are over-sampled by creating synthetic examples rather than duplication. The minority class is over-sampled by introducing synthetic examples along the line segments joining any/all of the k minority class adjacent neighbors. Depending upon the quantity of oversampling required, the neighbors from the k nearest neighbors are randomly chosen [19].

Proposed Cluster Concentric Circle Based Under-sampling (C3BUS) Method: The proposed approach in this work is different from the existing cluster based under sampling method [13]. This work mainly concentrates on choosing the instances in such a way that it does not miss any potentially useful instance while under sampling which is the main drawback of this sampling method. An imbalanced dataset is one where the distribution of samples among the classes are not the same which is in contrast to the balanced dataset. This approach is modelled using a sampling technique to build a balanced dataset from an imbalanced one. Different type of sampling methods include under sampling and over sampling. Under sampling could be either random under sampling or focussed under sampling. Here cluster based under sampling is considered and enhanced in such a way that the selection process would not miss the required sample. In this method, the dataset DS is divided into majority (DS_M) and minority (DS_m) samples.

$$DS = DS_M + DS_m \quad (1)$$

The majority samples are grouped into different k clusters using K means algorithm. This clustering algorithm is computationally faster than other hierarchical clustering algorithms [20].

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

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 ● 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).

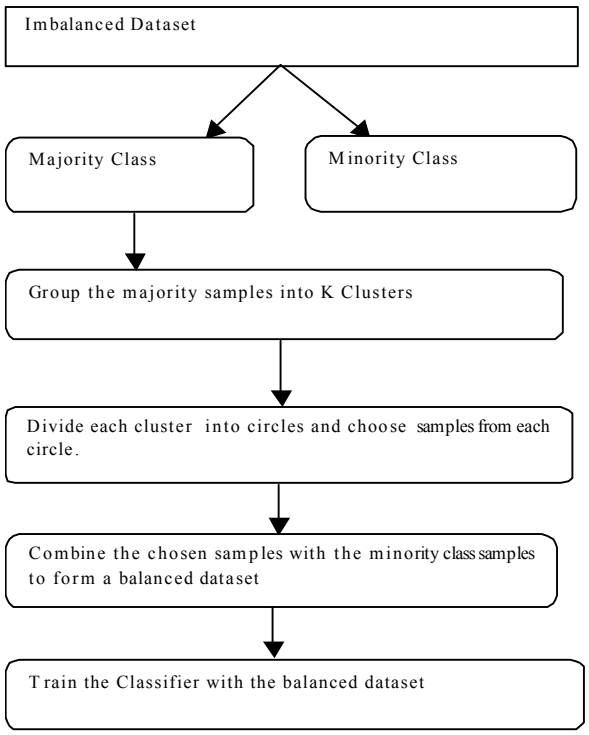

Fig. 2: Proposed C3BUS to build a balanced dataset

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

where np is the nearest sample in the cluster from the centroid. ☻ 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.

$$SC_{j+1} = \text{Max}\{d(SC_j, SCC_j)\}$$

$$[\text{where j= 1 to } nc_i] \quad (7)$$

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.

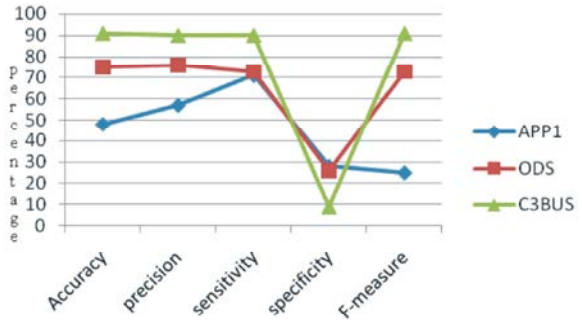

Fig. 3: Comparison of measures on Synthetic dataset

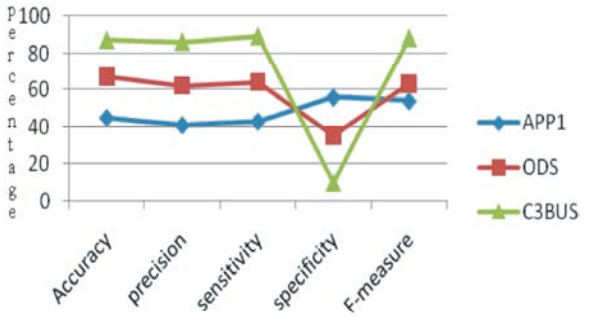

Fig. 4: Comparison of measures on Abalone Dataset

Table I: Performance of NN on Synthetic Dataset

`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

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

$$BDS = Ds_m (SC_j) \quad (8)$$

Experimental Analysis: This section evaluates the performance of the proposed C3BUS approach on synthetic datasets and also the datasets from the repository. Neural Network classifier is used to compare the performance of the C3BUS with the existing cluster based under-sampling approaches. Many research [21, 22] stated that Overall accuracy is not the only appropriate evaluation metric for a classifier with an imbalanced dataset. Since it does not consider the distribution of samples between the classes. In this work, five measures namely 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 used to evaluate the performance of the C3BUS, existing Cluster Based Under-sampling approach (APP1) and with original imbalanced dataset (ODS). In addition to these time is also considered. The proposed C3BUS is tested against a synthetic dataset, Abalone, Bioassay, Glass and Ecoli datasets with Neural Network classifier in MATLAB environment. It is inferred that the imbalance ratio has an influence in the performance of the classifier. The result fluctuates when there is a change in K value in K-means algorithm. The number of samples in the underrepresented class has a vital role in balancing the dataset. When there is very minimum number of samples in minority class it becomes tedious to balance the dataset. The time taken to classify the dataset with the proposed method is much lesser than the execution time with the existing method.

Results on Synthetic Dataset: The performance of C3BUS is evaluated by creating a synthetic dataset [23] with an imbalance ratio of 10:1. This dataset consist of 11000 instances (10000 negative samples and 1000 positive samples) and 5 attributes. It is found that the performance of C3BUS is better when compared with the existing method and Original imbalanced dataset. Table I and Fig. 3 depicts the results on Synthetic datasets.

Results on Abalone Dataset: To further test the performance of C3BUS Abalone dataset is used from KEEL repository with an imbalance ratio of 128:1. This dataset consist of 4174 instances with 9 attributes. Consistent with the results on Abalone dataset in this experiment, proposed method reports better performance. Table II and Fig. 4 pictures the performance.

Results on Bioassay Dataset: The performance of C3BUS is measured using the Bioassay dataset collected from UCI repository consisting of 3441 majority instances and 60 minority instances with 145 attributes. The imbalance ratio in this dataset is 57:1. Consistent with the results on this dataset in this experiment, proposed method achieve better performance. Table III and Fig 5 prove the result on Bioassay dataset.

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.

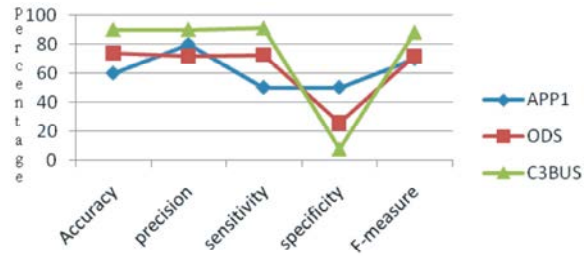

Fig. 5: Comparison of measures on Bioassay Dataset

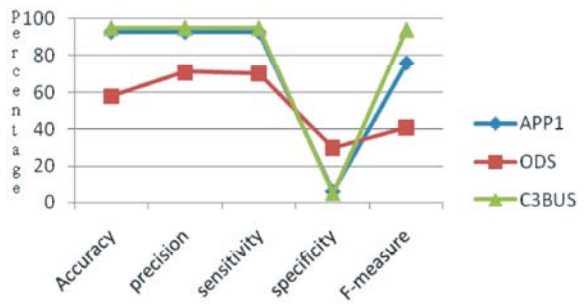

Fig. 6: Comparison of measures on Glass Dataset

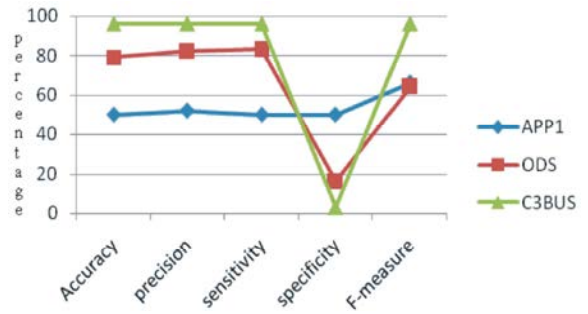

Fig. 7: Comparison of measures on Ecoli Dataset

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

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

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 C3BUSa

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

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.

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

CONCLUSION

When the class distribution is skewed it becomes a challenge for the classifier to correctly classify the underrepresented class. Several techniques have been proposed to alleviate the class imbalance problem. This work proposed a Cluster Concentric Circle based under-sampling (C3BUS) method to balance the imbalanced dataset. Later, the balanced dataset is given as an input to the neural network classifier to classify the samples. This work is compared with another Cluster based under-sampling method and proved to be better than the existing work in terms of Accuracy, precision, sensitivity specificity, F-Measure and execution time for the chosen datasets. The proposed method can be extended further by balancing the datasets with a hybrid sampling approach.

REFERENCES

1. Haibo He, 2009. Learning from Imbalanced Data. IEEE Transactions on Knowledge and data Engineering, 21(9).
2. Mikel Galar, A. Fernandez, E. Barrenechea and H. Bustince, 2011. 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, 42(4): 463-484.
3. Chawla, N.V., A. Lazarevic, L.O. Hall and K.W. Bowyer, 2003. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. In the Proceedings of. Seventh European Conf. Principles and Practice of Knowledge Discovery in Databases, pp. 107-119.

4. Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse and Amri Napolitano, 2010. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. IEEE transactions on systems, man and cybernetics—part a: Systems and Humans, 40(1).
5. Victoria Lopez, A. Fernandez, S. Garcia, V. Palade and F. Herrera, 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Elsevier Inc, 250: 113-141.
6. Sun, Y., M.S. Kamel, A.K.C. Wong and Y. Wang, 2007. Cost-Sensitive Boosting for Classification of Imbalanced Data. Pattern Recognition , 40: 3358-378.
7. Yubin Park and Joydeep Ghosh, 2014. Ensembles of alphaTrees for Imbalanced Classification Problems. IEEE transactions on knowledge and data engineering, 26(1): 131-143.
8. Yan-ping, Z., Z. Li-Na and W. Yong-Cheng, 2010. Cluster based majority under-sampling approaches for class imbalance learning. In the Proceedings of 2nd IEEE International Conference on Information and Financial Engineering (ICIFE), pp: 400-404.
9. Mollineda, V.G.J.S.R. and R.A.J. Sotoca, 2007. The class Imbalance problem in pattern classification and learning. Retrieved from http:// citeseerx.ist.psu.edu/ viewdoc/ download?doi=10.1.1.329.4200&rep=rep1&type=pdf.
10. Tom Fawcett, 2006. An introduction to ROC Analysis. Elsevier, Pattern Recognition Letters, 27(8): 861-874.
11. Yang, Q. and X. Wu, 2006. 10 challenging problems in data mining research. Int. Journal of Information Technology and Decision making, 5(4): 597-604.
12. Poolsawad, N., C. Kambhampati and J.G.F. Cleland, 2014. Balancing Class for Performance of Classification with a Clinical Dataset.In the Proceedings of the World Congress on Engineering 2014 Vol I, July 2 - 4, 2014, London, U.K.
13. Mostafizur Rahman. M. and D. N. Davis. Cluster based undersampling for unbalanced Cardivascular data. In the Proceedings of the world congress on Engineering, 2013 Vol III, WCE 2013, July 3-5, 2013.
14. Show-Jane Yen and Yue-Shi Lee, 2009. Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications. 36(3): 5718-5727.
15. Parinaz Sobhani, Herna Viktor and Stan Matwin, 2015. Learning from Imbalanced Data Using Ensemble Methods and Cluster-based Undersampling. New Frontiers in Mining Complex Patterns Lecture Notes in Computer Science 8983: 69-83.
16 Mr. Rushi Longadge, Ms. Snehlata S. Dongre, Dr. Latesh Malik, 2013. Multi-Cluster Based Approach for skewed Data in Data Mining. IOSR Journal of Computer Engineering (IOSR-JCE), pp: 66-73.
17 Kai-Biao Lin, Wei Weng, Robert K. Lai and Ping Lu, 2014. Imbalance Data Classification Algorithm based on SVM and Clustering Function in the 9th International Conference on Computer Science & Education (ICCSE 2014), Vancouver, Canada, pp: 544-548,.
18 Barendela, R., J.S. Sanchez, V. Garcia and E. Rangel, 2002. Strategies for Learning in class imbalance problems. Pattern Recognition, 36: 849-851.
19 Chawla, N., K. Bowyer, L. Hall and W. Kegelmeyer, 2002. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16: 321-357.
20 .http://playwidtech.blogspot.in/2013/02/k-means-clustering-advantages-and.html.
21. F.J. Provost and T. Fawcett, 1997. Analysis and Visualization of Classifier Performance: Comparision under Imprecise class and cost Distributions. Proc, Int'l Conf. Knowledge Discovery and Data mining, pp: 43-48.
22. Provost, F.J. T. Fawcett and R. Kohavi, 1998. The Case against Accuracy Estimation for Comparing Induction Algorithms. In the Proceedings of Int'l Conf. Machine Learning, pp: 445-453.
23. Bagyalakshmi, R, J. Malathi, K. Prathiba, Y. Samson, R Ravichandran and HN. Madhavan, 2012. A correlative study on Hepatitis C Virus Load Determined by Real Time Polymerase chain reaction with serum Biomarkers in patients with Renal disease. J. Mol. Biomark Diagn., 3(2).