Asian Journal of Information Technology 15 (16): 2887-2896, 2016
ISSN: 1682-3915

Hierarchical Sampling Techniques for Imbalanced Datasets

[1]S. Lavanya and [2]S. Palaniswami
[1]Department of CSE, Anna University Regional Campus, Coimbatore, Tamil Nadu, India
[2]Government College of Engineering, Bodiyanayakkanur, Tamil Nadu, India

Abstract: The imbalanced learning theory proposes varied distribution of data samples among different classes. According to this theory most of the samples get grouped under some classes and rest of the samples belong to the remaining classes. The solution for the problem can be provided by synthetic oversampling methods such as Majority Weighted Minority Oversampling Technique (MWMOTE). This method produces the artificial samples from the biased instructive alternative class samples by means of a clustering approach. Average-linkage agglomerative clustering is used to form clusters. The agglomerative clustering is not appropriate for large databases and has time complexity and high sensitive to noise. The proposed system introduces a clustering algorithm to adopt even for large database. Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is used in the proposed system. BIRCH algorithm clusters incoming multi-dimensional metric dataset and produces the unsurpassed clustering with the available resources dynamically. Another approach called Random Under Sampling (RUS) decreases the number of majority class dataset by randomly eliminating majority class data points currently in the training data set. The approach of using oversampling and under sampling is called the Re-sampling Technique. The performance comparison between the two methods is performed with the 14 data sets taken from the UCI repository. Experimental result exposes that the proposed system is competent in time complexity and providing high quality.

Key words: Imbalanced learning, under-sampling, oversampling, clustering, India

INTRODUCTION

In an imbalanced data set the classification categories are not represented equally (He and Garcia, 2009). Since, the class distributions among the classes are in deviating order, the classification problems of the imbalanced data sets are always significant. The data samples are unequally distributed among different classes in the problems based on imbalanced learning (Batista *et al*., 2004). One class may contain most of the samples and the other classes may contain the rest of the samples. Basically two classes are available, the majority classes and the minority classes (Fawcett and Provost, 1997). In an imbalanced data set many samples from one class are available while compared with the other classes. When minimum one class could be represented by a smaller number of training examples then the other classes forms the majority (Kubat *et al*., 1998; Ling and Li, 1998). Basically, classifiers will have high accuracy when dealing with majority class but very low accuracy on the minority classes. The reason behind this is the traditional method of training of the larger majority classes. In the present scenario many real-time problems dealing with machine learning are classified by imbalanced learning data where a minimum of one class is under represented while compared to others (Japkowicz *et al*., 1995; Clearwater and Stern, 1991; Japkowicz and Stephen, 2002). Since, the size of the data is unbounded and the nature is imbalanced, the data classification is very difficult and so the class imbalanced problem in data mining is a very big issue. To overcome the classification error classifier issues. But it is very difficult for a classifier to understand the class samples of the minority classes. Hence the class imbalanced problem is a great challenge (Weiss, 2004; Holte *et al*., 1989).

In this study, researchers propose a comparative analysis of a standard oversampling method Majority Weighted Minority Oversampling Technique (MWMOTE) with random under-sampling technique called Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH). The core of the analysis is to compare the performance metrics of the two methods namely accuracy, precision, recall, F-mean and G-mean.

Literature review: In this study, researchers compare two sampling methods, the first one MWMOTE which is an oversampling technique and the other BIRCH, an under-sampling technique. Random under sampling

Corresponding Author: S. Lavanya, Department of CSE, Anna University, Regional Campus, Coimbatore, Tamil Nadu, India

reduces the majority class data points currently in the training set. Like random oversampling, random under sampling has empirically performed well despite its simplicity (Tomek, 1976; Kubat and Matwin, 1997). Artificial over-sampling methods have been revealed to be extremely successful in addressing with imbalance dataset. ADASYN (He *et al*., 2008) is a technique to generate minority data samples based on their distributions. This techniques can reduce the learning bias formed by the actual imbalanced data distribution and also can shift the decision boundary to focus on the samples which are difficult to learn (Wu *et al*., 2007).

RAMOBoost (Chen *et al*., 2010; Cieslak and Chawla, 2008; Freund and Schapire, 1996) is another oversampling method for assigning weights adaptively the minority class samples. Since if weight is large, many synthetic samples could be generated from the corresponding minority class samples. Borderline SMOTE (Han *et al*., 2005) is an oversample method to recognize the seed samples which are termed as minority class samples of the border line. This technique utilizes the seed samples to generate synthetic samples in the boarderly neighbourhood (Freund and Schapire, 1997).

MATERIALS AND METHODS

The above architecture diagram (Fig. 1) explains the initial stage of operation starts from collection of data sets from UCI repository. The data sets is pre-processed, the classification of minority and majority sets done. Then two parts of operations are performed, in the part one the minority sets are processed and in the part two the majority sets are processed. In the part one initially the minority sets are chosen and noise minority class samples are removed (Lewis and Catlett, 1994). Then, the weight value is computed, after this computation the closed factor and density factors are computed. In the part two initially the majority sets are chosen and from that, majority class are selected randomly (Liu *et al*., 2006). Then the selected samples are deleted followed by computation of synthetic is done. Now, balanced data is obtained from both part one and part two (Wang and Yao, 2012). This data is fed to the classification section and the performance evaluation is done. Majority Weighted Minority Over sampling Technique (MWMOTE) algorithm can be found in.

Random under sampling: The re-sampling process is done by Random under sampling method. The ratio between the majority and minority sample is the specified level for eliminating random values in the training set of majority class. Random under sampling causes the majority class details to be omitted, such as decision boundary between majority and minority classes (Liu *et al*., 2009).

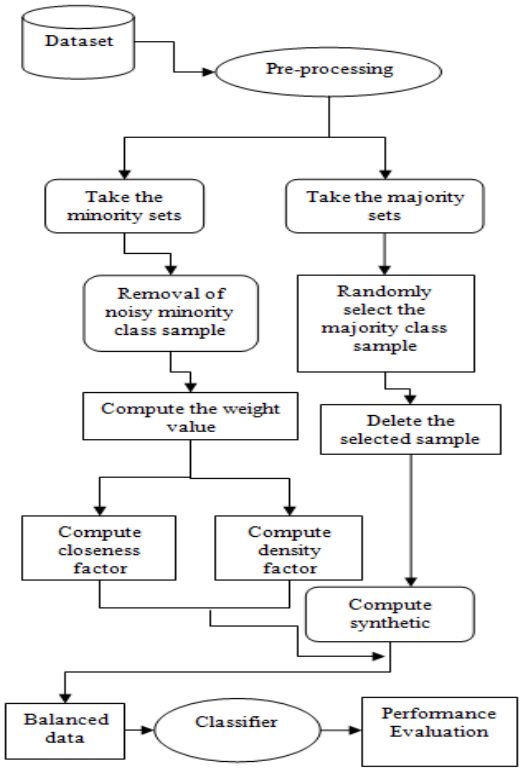

Fig. 1: Architecture diagram for proposed sysytem

The four re-sampling methods introduced here are NearMiss1, NearMiss2, NearMiss3 and Distant1. In the first method the majority class samples in the training set is the smallest. In the second method the standard distance from the three furthest members are considered to be the smallest. In NearMiss3, the n closest majority class document in the training set is considered.

Generating the synthetic samples: The realization of MWMOTE is basically dependent on how division the set S_{min}. MWMOTE (Barua *et al*., 2014) uses average-linkage agglomerative gathering, a hierarchical clustering procedure for the above reason. Here, clusters are formed in a bottom-up tactic. The steps are stated below (Assume D data samples are given as input):

- Step 1: Allocate each sample to individual cluster, at the beginning there will be D clusters of size one
- Step 2: Locate two neighbouring clusters state $L_i$and L_j

- Step 3: Combine the clusters L_i and L_j to form acluster L_m
- Step 4: Revise the distance measures among the recently joined cluster and all the former cluster(s)
- Step 5: Replication of Steps 2-4 can be carried out to form a single cluster of size D

The algorithm mentioned above creates one cluster of size D. In case of forming n clusters, the algorithm can be terminated in step 3 at any stage. MWMOTE make use of a threshold, T_h and terminates the merging procedure when the distance among neighbouring pair exceeds T_h. Then, the yield will be the rest of clusters enduring at that point. This algorithm may produce an assorted number of clusters for the similar type of data where the only dissimilarity is the dimension of the eminence space. The second difficulty of using a persistent T_h is in few datasets, the samples are rather sparse while in rest of the sets, they are dense. A stable T_h will form few clusters for the datasets where the mean distance is small and remaining clusters have large mean distance. The perceptible point here is that T_h must be the data reliant and it should be computed using some heuristics methods to measure the distance between data samples. In this work, calculate T_h as follows (Barua *et al*., 2014). Determine the standard distance d_{avg} as:

$$d_{avg} = \frac{1}{|S_{min\,f}|} \sum_{x \in Smin\,f} \min_{y \neq x, y \in} S_{min\,f}\{dist(x,y)\} \quad (1)$$

Compute T_h as the product of d_{avg} and a constant parameter, C_p as:

$$T_h = d_{avg} \times C_p \quad (2)$$

Determine the least Euclidean distance to any other member in the similar set for every member of C_p. Then, calculate the mean of all these distances to find C_p. The constant C_p modifies the output of the clustering technique. The increase in C_p will upturn the cluster size but lessens the number of clusters. Hence, the vice-versa.

Birch clustering: BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) belonging to unsupervised data mining algorithm performs hierarchical clustering over a very large dataset. BIRCH has an eminent advantage in its ability to form progressive and dynamic cluster from the incoming multi-dimensional data points to produce the high quality clustering for any dataset (memory and time constraints).

BIRCH is the first clustering algorithm proposed in the data handling techniques to handle 'noise'. The existing clustering algorithms shows less performance over a large databases and not in consideration because a large data-set needs more space in main memory to fit in. As a result, overhead was high.

Consider a set of N d-dimensional data points, the Clustering Feature (CF) of the set is defined as the triple CF = {N, LS, SS} where Linear Sum (LS) and Square Sum (SS) of data points. Clustering features are arranged in a CF tree, a height balanced tree with branching factor B and threshold T as two parameters. Each non-leaf node contains at most B entries consisting of $child_i$, a child node pointer and $child_i$, the associated sub cluster's clustering feature represented as CF, $child_i$. At most L entries each of the form $[CF_i]$ are available in individual leaf node. All leaf nodes are chained by its pointers 'prev' and 'next' and the size of the tree are governed by the parameter T. A page size P is decided to fit a node in memory. B and L resolute P. So, P can be varied for tuning the performance (Quinlan, 1986).

RESULTS AND DISCUSSION

Experimental study: The performance of the MWMOTE is compared using agglomerative clustering with the proposed research. We use single datasets with dissimilar difficulty and imbalance percentage. We also utilize 14 datasets composed from the UCI machine-learning database (Murphy and Aha, 1994). Two methods are performed on these data sets: MWMOTE with K-means clustering and Both MWMOTE and random under sampling with BIRCH clustering. The performance evaluation is based on the comparison of these two methods.

Experiments on real-world data sets: This section presents the performance evaluation of MWMOTE (Barua *et al*., 2014) and k-means clustering (Mani and Zhang, 2003) and BIRCH clustering. The parameters involved in comparison of clustering algorithms are as follows.

Closeness factor; C_f (y_i, x_i): Primarily, calculate the normalized Euclidean distance:

$$d_n(y_i, x_i) = \frac{dist(y_i, x_i)}{1} \quad (3)$$

Table 1: Selection weight and probability for two methods

	Using closeness factor alone		Using both density factor and closeness factor	
Data sets	Selection weight	Selection probability	Selection weight	Selection probability
Breast cancer	12360	1.55787	14526	1.57732
Breast- tissue	5139	1.21582	5241	1.41378
CTG	6490	0.98625	6672	0.97437
Glass	8542	2.07532	8684	2.14678
Libra	14340	0.98620	14674	1.09632
Pima	8346	1.57210	8472	1.74681
Robot	3078	0.98563	3370	1.07328
Urban land cover	5604	2.98542	5965	3.01647
Vehicle	8947	0.86348	9045	0.97562
Yeast	11276	1.00126	12643	1.76348
Abalone	2850	0.7855	3076	1.17532
Ecoli	3386	1.76402	4174	1.76846
Pageblock	10674	2.06516	12042	2.18549
Wine	8466	2.15783	9056	2.55925

Where:

dist (y_i, x_i) = The Euclidean distance from y_i to x_i

l = The measurement of the feature space

Then, calculate $C_f(y_i, x_i)$ in the subsequent way:

$$C_f(y_i, x_i) = \frac{f\left(\frac{1}{d_n(y_i x_i)}\right) \times CMAX}{C_f(th)} \quad (4)$$

wherever, C_f (y_i, x_i) and CMAX are the user defined limitation and f is a cut-off task.

Density factor: The density factor D_f (y_i, x_i) explain that the light cluster must have extraartificial samples than the intense cluster when the group are equidistant from the resolution boundary.

$$D_f(y_i, x_i) = \frac{C_f(y_i x_i)}{\sum_{q \in S_{i\min}} C_f(y_i x_i)} \quad (5)$$

Comparison between MWMOTE and re-sampling base on size: Table 1 explain the method of selecting weight and probability for two methods. Table 2 gives an overall comparison number of samples when performing oversampling Table 3 represents provides the results of oversampling and under sampling based on the label size. Table 4 explain a confusion matrix of the two classes along with its characteristics.

The relation between the actual data and predicted data s provided by confusion matrix. Under-sampling is an efficient class-imbalance learning method that uses only a subset of major class. The main discrepancy of under-sampling is that many major class examples are overlooked. In order to overcome this disadvantage, two algorithms are proposed. To measure the classifier performance, cumulative sum of True Positive (TP), False Positive (FP), True Negative (TN), False Negative

Table 2: No. of labels when performing only oversampling

Data sets	Actual minority size	Actual majority size	Final oversampling size
Breast cancer	47	151	151
Breast- tissue	14	22	22
CTG	19	285	285
Glass	7	23	23
Libra	11	42	42
Pima	39	71	71
Robot	328	2205	2205
Urban land cover	29	122	122
Vehicle	199	218	218
Yeast	5	463	463
Abalone	6	98	98
Ecoli	3	33	33
Pageblock	33	749	749
Wine	7	19	19

Table 3: No. of labels when performing both oversampling and under sampling

Data sets	Actual minority size	Actual majority size	Final oversampling size	Final undersampling size
Breast cancer	47	151	99	99
Breast- tissue	14	22	18	18
CTG	19	285	152	152
Glass	7	23	15	15
Libra	11	42	27	27
Pima	39	71	55	55
Robot	328	2205	1266	1266
Urban land cover	29	122	75	75
Vehicle	199	218	208	208
Yeast	5	463	234	234
Abalone	6	98	52	52
Ecoli	3	33	18	18
Pageblock	33	749	391	391
Wine	7	19	13	13

Table 4: Confusion matrix

	True class	
Data set	Positive	Negative
Positive	TP	FP
Negative	FN	TN
Row Sum	P	N

Table 5: Performance measure of MWMOTE with k-means clustering

Data sets	Accuracy	Precision	Re-call	F-mean	G-mean
Breast cancer	88.5417	0.8450	0.8450	0.8450	0.9260
Breast-tissue	91.5094	0.9156	0.9179	0.9168	13.0402
CTG	90.4547	0.9122	0.9034	0.8956	0.9196
Glass	86.3476	0.8872	0.8958	0.8450	0.8754
Libra	89.7634	0.9014	0.8356	0.9042	1.8965
Pima	84.8214	0.8652	0.8754	0.8826	2.7632
Robot	87.0449	0.8075	0.8616	0.8337	0.0214
Urban land cover	89.9259	0.8674	0.9047	0.8857	2.1175
Vehicle	90.9846	0.9124	0.9094	0.9114	1.0879
Yeast	89.8042	0.7352	0.9039	0.8109	0.0038
Abalone	84.1602	0.8347	0.7834	0.8614	0.0478
Ecoli	90.1672	0.9418	0.9016	0.9326	1.0238
Pageblock	97.4618	0.9262	0.9418	0.9628	0.8416
Wine	84.7842	0.8653	0.8412	0.8672	1.2532

Table 6: Actual data and predicted data calculation using confusion matrix

		MWMOTE with k-means clustering				Re-sampling with birch clustering			
Data sets	Class label	True positive	False positive	False negative	True negative	True positive	False positive	False negative	True negative
Breast cancer	‘1’	134.00	11.00	11.00	36.00	140.00	1.00	5.00	46.00
	‘2’	36.00	11.00	11.00	134.00	46.00	5.00	1.00	140.00
Breast- tissue	‘fad’	20.00	2.00	2.00	82.00	21.00	1.00	1.00	84.00
	‘adi’	18.00	0.00	3.00	85.00	20.00	1.00	1.00	84.00
CTG	‘3’	220.00	45.00	32.00	136.00	226.00	38.00	24.00	145.00
	‘6’	158.00	22.00	18.00	98.00	164.00	13.00	11.00	112.00
Glass	‘5’	54.00	16.00	24.00	48.00	55.00	14.00	8.00	62.00
	Others	32.00	7.00	7.00	28.00	38.00	3.00	1.00	39.00
Libra	‘2’	414.00	7.00	46.00	1014.00	414.00	7.00	46.00	1014.00
	‘3’	5.00	19.00	0.00	457.00	5.00	19.00	0.00	457.00
Pima	‘1’	320.00	62.00	56.00	288.00	342.00	46.00	28.00	310.00
	‘0’	112.00	49.00	55.00	148.00	136.00	37.00	41.00	164.00
Robot	‘left’	1915.00	181.00	245.00	3070.00	2047.00	84.00	128.00	3167.00
	‘right’	1799.00	178.00	244.00	98.00	1952.00	81.00	145.00	3248.00
Urban land cover	‘2’	55.00	6.00	4.00	610.00	58.00	1.00	4.00	552.00
	‘1’	109.00	1.00	1.00	615.00	118.00	1.00	4.00	552.00
Vehicle	‘1’	194.00	14.00	21.00	614.00	414.00	7.00	46.00	1014.00
	Others	202.00	28.00	15.00	602.00	5.00	19.00	0.00	457.00
Yeast	‘9’	414.00	7.00	46.00	1014.00	428.00	4.00	32.00	1017.00
	‘7’	5.00	19.00	0.00	457.00	5.00	9.00	0.00	1467.00
Abalone	‘18’	143.00	24.00	31.00	264.00	156.00	14.00	22.00	288.00
	‘9’	56.00	18.00	12.00	88.00	73.00	11.00	4.00	104.00
Ecoli	‘0’	220.00	45.00	32.00	136.00	226.00	38.00	24.00	145.00
	‘1’	158.00	22.00	18.00	98.00	164.00	13.00	11.00	112.00
Pageblock	‘graphics’	20.00	2.00	2.00	82.00	21.00	1.00	1.00	84.00
	Others	18.00	0.00	3.00	85.00	20.00	1.00	1.00	84.00
Wine	‘3’	2047.00	84.00	128.00	3167.00	2089.00	57.00	112.00	3324.00
	Others	1952.00	81.00	145.00	3248.00	2068.00	67.00	123.00	3456.00

(FN) mode are calculated and develop a confusion matrix. Various performance measures considered for evaluation can be defined as:

$$\text{Accuracy} = \frac{TP + TN}{p + n}$$

$$\text{False Positive Rate(FPR)} = \frac{FP}{FP + TN}$$

$$\text{True positive rate}(ACC_+) = \frac{TP}{TN + FP}$$

$$\text{True positive rate}(ACC_-) = \frac{TN}{TN + FP}$$

$$G - \text{Mean} = \sqrt{ACC_+} \times ACC_-$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FP} = ACC_+$$

$$F - \text{Measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Comparison with two clustering methods: Table 5 presence the overall performance measure of MWMOTE with K-means clustering. Table 6 and 7 provides performance metric of two methods, i.e., MWMOTE with k-means clustering and Both MWMOTE and random under sampling with BIRCH clustering. Our MWMOTE,

Table 7: Performance measure of MWMOTE and random under sampling with BIRCH clustering

Data sets	Accuracy	Precision	Re-call	F-mean	G-mean
Breast cancer	96.8750	0.9474	0.9721	0.9596	0.8947
Breast-tissue	94.3396	0.9416	0.9410	0.9413	13.1293
CTG	93.4578	0.9256	0.9412	0.9274	0.9248
Glass	90.3286	0.9143	0.9122	0.9362	0.9412
Libra	97.6012	0.9214	0.8672	0.9148	2.0146
Pima	88.3452	0.8732	0.9214	0.9176	2.9014
Robot	97.0449	0.9412	0.8915	0.9037	0.0206
Urban land cover	97.6296	0.9696	0.9777	0.9736	2.1635
Vehicle	92.6453	0.9271	0.9264	0.9264	0.9267
Abalone	94.2606	0.8148	0.9525	0.8783	0.0044
Abalone	96.2678	0.9046	0.8842	0.8935	0.5614
Ecoli	96.5618	0.9216	0.8918	0.9156	1.0634
Pageblock	88.2759	0.8914	0.9216	0.9412	0.8682
Wine	90.1437	0.9056	0.9214	0.8856	1.1438

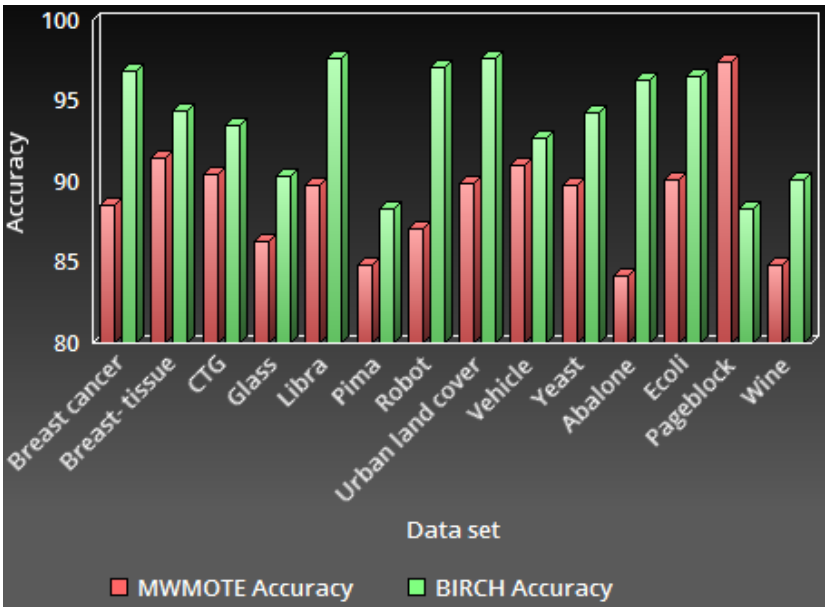

Fig. 2: Accuracy measure of MWMOTE and BIRCH clustering

however is pertaining both divisioning and oversampling simply on the alternative class samples.

Table 7 represents performance measure chart for the two methods. Some performance measures are used in imbalanced knowledge. They are precision, recall, F-measure and geometric-mean (G-mean). The performance evaluation is based on the comparison of these two methods. These examples are those that are frequently located close to the decision limit and go to the small-sized clusters.

Simulation results: Figure 2 explains comparison between datasets and accuracy. From the graph it's clear that for the datasets libra the accuracy percentage is comparatively higher.

Figure 3 explains comparison between datasets and Recall. From the graph it's clear that from the dataset Urban land cover the recall percent comparatively higher. Figure 4 given bellow explains comparison between datasets and Precision. From the graph it's clear that for the datasets Urban land cover the precision percentage is comparatively higher.

Figure 5 given bellow explains comparison between datasets and F-Mean. From the graph it's clear that for the datasets Urban land cover the F-Mean percentage is comparatively higher.

Figure 6 explains comparison between datasets and G-Mean. From the graph it's clear that for the datasets Breast cancer the G-Mean percentage is comparatively higher.

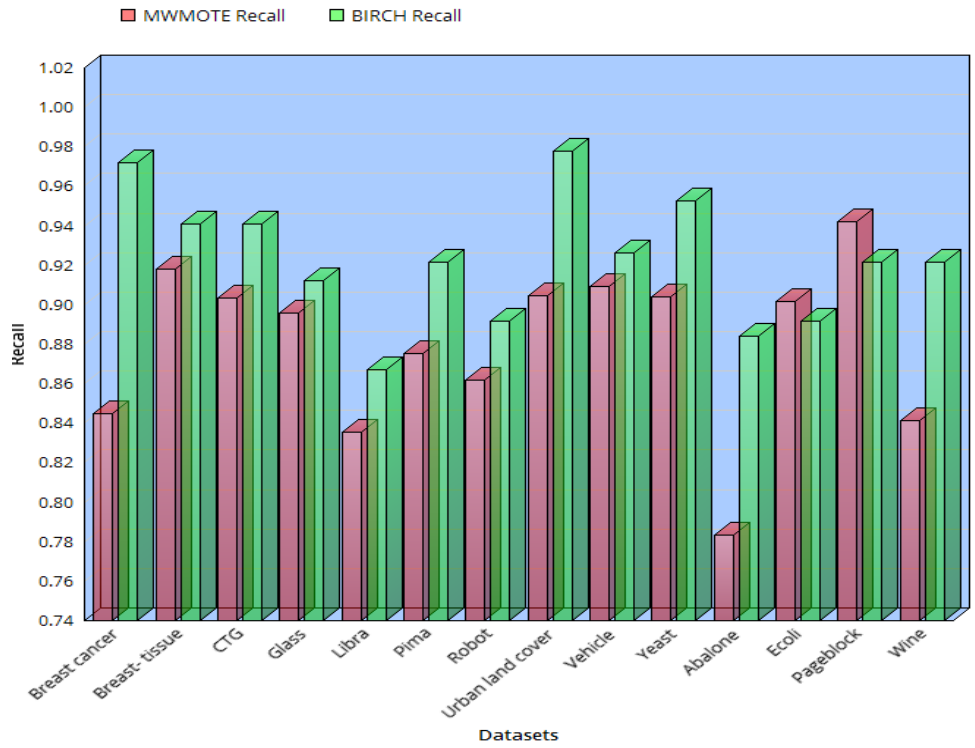

Fig. 3: Recall measure of MWMOTE and BIRCH clustering

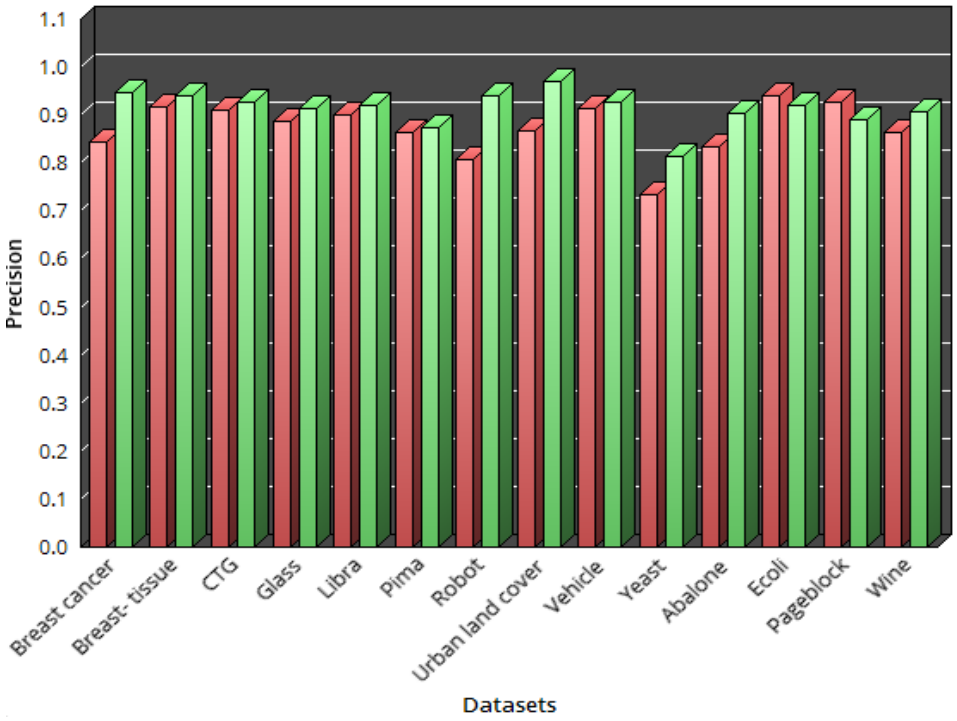

Fig. 4: Precision measure of MWMOTE and BIRCH clustering

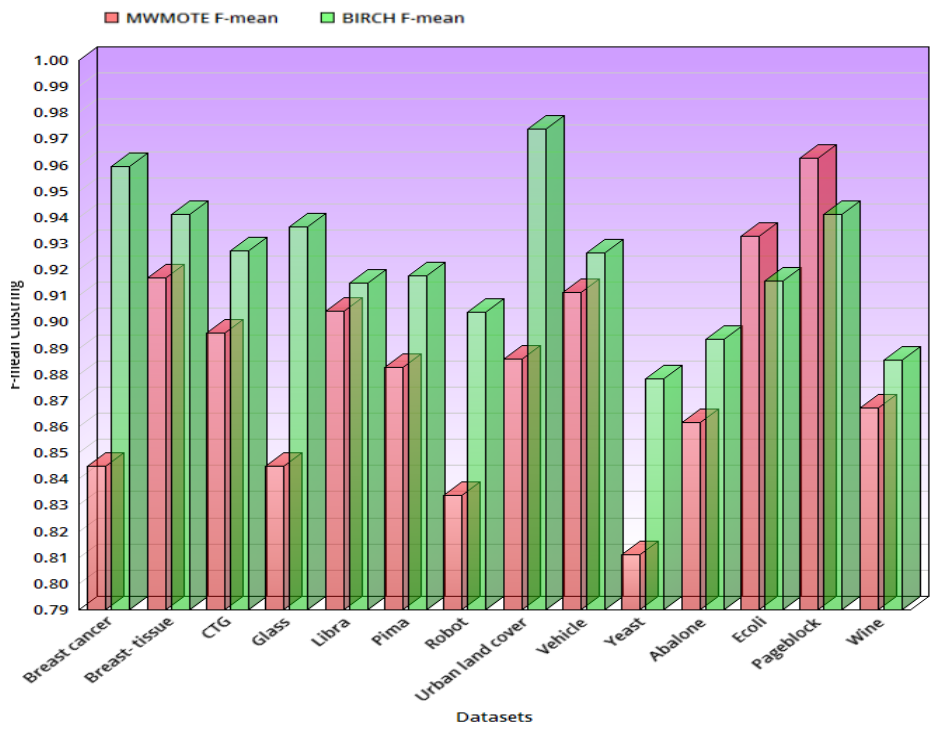

Fig. 5: F-Mean measure of MWMOTE and BIRCH clustering

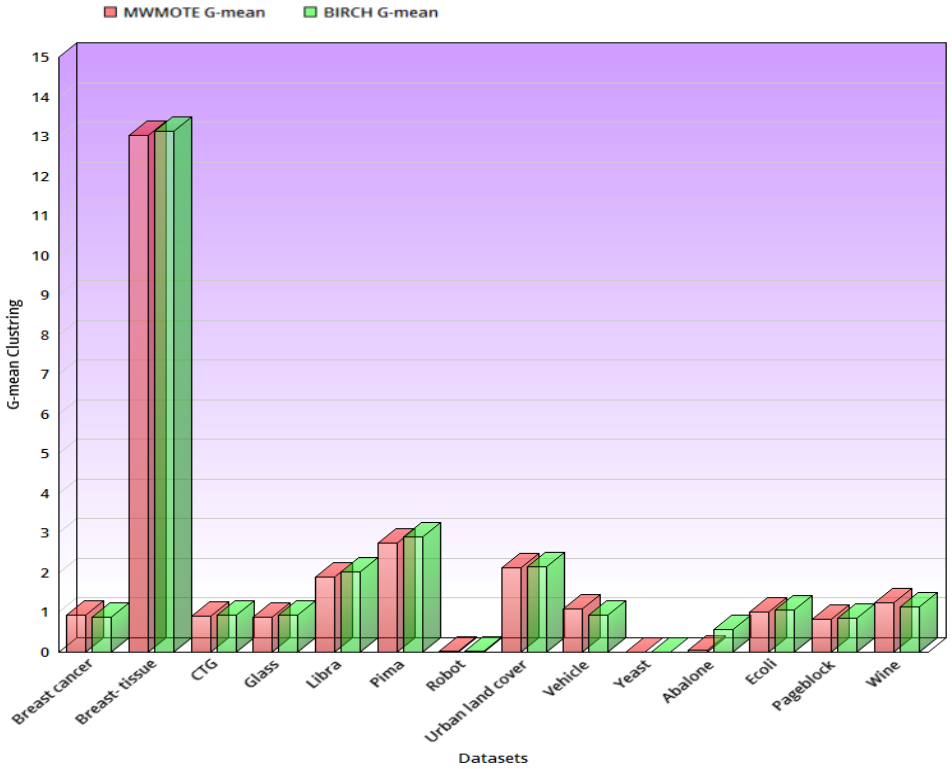

Fig. 6: G-Mean measure of MWMOTE and BIRCH clustering

CONCLUSION

The MWMOTE routines the majority class samples close to the decision boundary to effectively choose the minority class samples. The scope of gathering is to safeguard that the engendered samples should be within the minority class region for checking any incorrect or noisy simulated sample creation. The deficiency in the MWMOTE method is high time complexity and it is high subtle to noise. Whereas in the Birch clustering time complexity is less. BIRCH can provide good clustering with a scanned data and the quality can be improved further with a few additional scans. The procedure of using the resampling technique provides better classification method, i.e., both synthetic oversampling and random under sampling method and can be used over large datasets.

REFERENCES

Barua, S., M.M. Islam, X. Yao and K. Murase, 2014. MWMOTE-majority weighted minority oversampling technique for imbalanced data set learning. IEEE Trans. Knowledge Data Eng., 26: 405-425.

Batista, G.E.A.P.A., R.C. Prati and M.C. Monard, 2004. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsl., 6: 20-29.

Chen, S., H. He and E.A. Garcia, 2010. RAMOBoost: Ranked minority oversampling in boosting. IEEE Trans. Neural Networks, 21: 1624-1642.

Cieslak, D.A. and N.V. Chawla, 2008. Start globally, optimize locally, predict globally: Improving performance on imbalanced data. Proceedings of the 2008 8th IEEE International Conference on Data Mining, December 15-19, 2008, Pisa, pp: 143-152.

Clearwater, S.H. and E.G. Stern, 1991. A rule-learning program in high energy physics event classification. Comput. Phys. Comm., 67: 159-182.

Fawcett, T. and F. Provost, 1997. Adaptive fraud detection. Data Mining Knowledge Discovery, 3: 291-316.

Freund, Y. and R.E. Schapire, 1996. Experiments with a new boosting algorithm. Proceedings of the 13th International Conference on Machine Learning, July 3-6, 1996, Bari, Italy, pp: 148-156.

Freund, Y. and R.E. Schapire, 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci., 55: 119-139.

Han, H., W.Y. Wang and B.H. Mao, 2005. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. Proceedings of the International Conference on Intelligent Computing, August 23-26, 2005, Hefei, China, pp: 878-887.

He, H. and E.A. Garcia, 2009. Learning from imbalanced data. IEEE Trans. Knowledge Data Eng., 21: 1263-1284.

He, H., Y. Bai, E.A. Garcia and S. Li, 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. Proceedings of the IEEE International Joint Conference on Neural Networks, June 1-8, 2008, Hong Kong, pp: 1322-1328.

Holte, R.C., L.E. Acker and B.W. Porter, 1989. Concept learning and the problem of small disjuncts. Proceedings of the International Joint Conference on Articial Intelligence, July 16-22, 1989, Morgan Kaufmann, pp: 813-818.

Japkowicz, N. and S. Stephen, 2002. The class imbalance problem: A systematic study. Intell. Data Anal., 6: 429-449.

Japkowicz, N., C. Myers and M. Gluck, 1995. A novelty detection approach to classification. Proceedings of the 14th International Joint Conference on Artificial Intelligence, August 20-25, 1995, Canada, pp: 518-523.

Kubat, M. and S. Matwin, 1997. Addressing the curse of imbalanced training sets: One-sided selection. Proceedings of the 14th International Conference on Machine Learning, July 8-12, 1997, Nashville, November 18, 2016Tennessee, USA, pp: 179-186.

Kubat, M., R.C. Holte and S. Matwin, 1998. Machine learning for the detection of oil spills in satellite radar images. Mach. Learning, 30: 195-215.

Lewis, D.D. and J. Catlett, 1994. Heterogeneous uncertainty sampling for supervised learning. Proceedings of the 11th International Conference on Machine Learning, (ICML'94), Morgan Kaufmann, pp: 148-156.

Ling, C.X. and C. Li, 1998. Data mining for direct marketing: Problems and solutions. Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, August 27-31, 1998, New York, USA., pp: 73-79.

Liu, X.Y., J. Wu and Z.H. Zhou, 2006. Exploratory undersampling for class-imbalance learning. Proceedings of the IEEE International Conference on Data Mining, December 18-22, 2006, Hong Kong, pp: 965-969.

Liu, X.Y., J. Wu and Z.H. Zhou, 2009. Exploratory under sampling for class imbalance learning. Proceedings of the 15th International Conference on Auditory Display, May 18-21, 2009, Copenhagen, Denmark, pp: 919-926.

Mani, I. and I. Zhang, 2003. KNN approach to unbalanced data distributions: A case study involving information extraction. Proceedings of the Workshop on Learning from Imbalanced Datasets, (WLID'03), Washington, DC., USA -.

Murphy, P.M. and D.W. Aha, 1994. UCI repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine, CA.

Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn., 1: 81-106.

Tomek, I., 1976. Two modifications of CNN. IEEE Trans. Syst. Man Cybern., 6: 769-772.

Wang, S. and X. Yao, 2012. Multiclass imbalance problems: Analysis and potential solutions. IEEE Trans. Syst. Man Cybernet., 42: 1119-1130.

Weiss, G.M., 2004. Mining with rarity: A unifying framework. ACM SIGKDD Explorations Newsl., 6: 7-19.

Wu, J., H. Xiong, P. Wu and J. Chen, 2007. Local decomposition for rare class analysis. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 12-15, 2007, San Jose, CA., pp: 814-823.