https://doi.org/10.26438/ijcse/v6si1.19

IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-6, Special Issue-1, Feb 2018 E-ISSN: 2347-2693

Handling of Class Imbalanced Problem in Big Data Sets: An Experimental Evaluation (UCPMOT)

S.S. Patil[1*], S. P. Sonavane[2]

[1*]CSE, Rajarambapu Institute of Technology, India
[2] IT, Walchand College of Engineering, India

[*]*Corresponding Author: sachin.patil@ritindia.edu, Tel.: +91-997-070-0925*

Available online at: www.ijcseonline.org

***Abstract*—** The huge amount of NoSQL data has acknowledged a new provision of context for processing. A new trail of data handling technologies with massive resources assists to store and process these gigantic data sets. The current attention is to determine the undisclosed information by assimilating this data bulks & handling it as per use. Further they are pre-processed and converted for needful analysis. The volume and variety of these data sets endure rising relentlessly. Moreover, imbalanced in many real-worlds vast data sets have elevated a point of concern in the research domain. The skewed distribution of classes in the data sets poses a difficulty to learn using traditional classifiers. They tend more towards majority classes. In recent years, numerous solutions have been proposed to address imbalanced classification. However, they fail to address the various data characteristics such as overlapping, redundancy involving classification performance. A rational over_sampling technique i.e. Updated Class Purity Maximization Over_Sampling Technique using Safe-Level based synthetic sample creation is proposed to efficiently handle imbalanced data sets. The newly suggested Lowest versus Highest method addresses the handling of multi-class data sets. The data sets from the UCI repository are processed using the mapreduce based programming on Hadoop framework. The evaluation parameters viz. F-measure and AUC are used to authenticate the performance of proposed technique over benchmarking techniques. The results attained evidently quote the dominance of the proposed technique.

***Keywords*—**Imbalanced datasets, Big Data, Over_sampling techniques, Multi-class, Safe-Level based Synthetic Samples

I. HANDLING OF IMBALANCED DATASETS:

INTRODUCTION

The data in the form of massive volume, extreme velocity and varied variety has lead to today's catchphrase 'Big Data'. The challenges set by the Big Data analytics are to be addressed capably. Huge digital Big Data including its varied forms evolving per day has outdrawn the need of cutting-edge analytics. In addition there is a requirement to exploit the streaming data with the capable conduct of analysis.

The superior verdict prediction of the inferred information from the massive diverse data is a challenge [1]. The volume of data is estimated to increase by 20 times than the current date [2-3]. To deal with the challenges evolved in Big Data management has set a crucial inclination [3-4]. Furthermore, the capability of the ecosystem to deal with usage, mobility and deployment of data has to be emphasized [5-6].

Classification of the minority samples appropriately in imbalance scenario has become the main focus of study [11]. Generally the classifiers ignore the minority instances while forming rule sets. The numerous real-world applications are affected by class imbalance problem wherein the number of samples in one class is very marginal compared to other classes [7~9, 34]. Issues in fields related to software defect detection [10], threat supervision, medical judgment, web author identification [36] and similar have drawn attention towards concerns of multi-class imbalanced data sets. The representation of boundaries in imbalanced data sets is a difficult concern for learning algorithms. Skewed data partition is an integral issue for learning of classifiers.

Updated Class Purity Maximization Over_Sampling Technique (UCPMOT) is a superior over_sampling technique presented in this paper. It acquaints the class imbalance problem. The basic over_sampling process using safe-level based displacement factor is carried out with the help of other two over_sampling techniques (Non-cluster/Cluster based). The experiments are conducted on Hadoop framework using the distributed mapreduce structure [14-15]. Two classifiers viz. Random Forest and MultiLayer Perceptron [12-13] are used to perform classification. The preciseness of techniques is assessed by using two measures: F-measure and AUC values.

Remaining of the paper is organized as follows, Section I contains the introduction of imbalanced data sets, Section II contain the related work, Section III comprise the organization of the proposed work, Section IV comprehend the details of projected technique, Section V outline the experimental settings, Section VI describes the experimental

evaluation, Section VII concludes the research work with future directions.

II. RELATED WORK

Classification of imbalanced data sets is recognized by numerous available techniques working at dissimilar levels. They are broadly considered into three levels viz. data level, procedure level and cost-sensitive level [11, 14]. Data level works by updating the size of the data sets. The predominant techniques at procedure level work with the processes to manage imbalanced Big Data sets. The cost–sensitive technique is a mix of both techniques viz. data level and procedure level. The techniques discussed in this paper deal with the data level technique. The data level technique is categorized into three types: Undersampling, Over_sampling and Hybrid technique [11, 14]. Over_sampling may incline to reproduce noisy data, whereas undersampling might lose the useful data. The easiest way to deal with under_over sampling is random approach [16]. Over_sampling results show extra advantages than the results of undersampling techniques. The recommended techniques work in alignment with the over_sampling approaches.

Synthetic Minority Oversampling Technique (SMOTE) algorithm [17] is one of the basic over_sampling techniques. It works on the class imbalance issue by synthesizing the minority class examples. 'K' Nearest Neighbors (KNN) are selected randomly to satisfy the over_sampling rate. SMOTE encounters some drawbacks including over-generalization and lack of systematizing disjuncts. Enhanced techniques such as Borderline-SMOTE [18], SafeLevel-SMOTE [19] and Adaptive Synthetic Sampling (ADASYN) [20] help to overcome these drawbacks. The proposed technique follows the same baseline while leveraging the disjuncts and generalization issue. Evolutionary algorithms resolve the imbalanced Big Data sets issue using the technique belonging to nested generalized model, considering objects in Euclidean n-space [21]. Boundary based oversampling technique used in SMOTE+GLMBoost and NRBoundary-SMOTE [22] are engaged to resolve imbalance data set problems. The UCPMOT technique assists to engage farthest borderline neighbors and their mean, involving the nearest samples. The ensemble techniques viz. SMOTEBoost [23], AdaBoost [24] and RUSBoost are tangled with SMOTE to work over the problems of the imbalanced data set. In [25], fuzzy rule classification is anticipated as a solution for the multi-class dilemma by merging the pairwise learning with preprocessing. Ultimately the LVH method clamps meritoriously the issues of over_sampling in multi-class imbalanced data sets. The ensemble based techniques (Random Forest) helps to effectually discourse classification analysis [26, 35]. They are validated as scalable, durable and capable of handling categorical data. In [27], an incremental clustering based fault detection approach is studied. This includes extreme class distributions of Gaussian/non-Gaussian types and process drifts. The ordinal classification of imbalanced Big Data sets in [28], approximates the class probability distribution using the weighted KNN technique. Competent string based procedure to detect class in data streams is reflected in [29]. It includes attributes of infinite-length, concept-evolution and data drift. The procedure to aid the valuation of domain samples methodically is proposed as Mega-Trend-Diffusion Technique (MTDF) in [40] to address the class imbalance problem. A recent imbalanced data set handling technique i.e. Majority Weighted Minority Oversampling Technique (MWMOTE) [41] efficiently recognizes those minority instances which are difficult in terms of learning. It assigns the weight to each of them based on Euclidean distance from the nearest majority class samples. The artificial samples are created from these samples using a clustering approach. The use of the immune network [42] coordinates the immune centroids as synthetic instances, based on high data density clusters which help to handle the imbalanced data sets. It implicitly encourages the broadening of the minority class decision space.

III. PROPOSED ORGANIZATION

The proposed architecture of experimental work [30] involves the analysis of the over_sampling effect on the imbalanced data set to enhance the classification results. The procedure involves to store, process and analyze the produced balanced data set. The over_sampling techniques are performed using Hadoop environment.

The projected over_sampling techniques (non-cluster and cluster based) works with binary as well as a nonbinary-class group of data sets. A newly suggested method Lowest versus Highest (LVH) [30] effectively mechanize the treatment of nonbinary-class data sets. The uppermost majority class is considered for over_sampling versus each of the positive class (satisfying imbalance ratio (I.R.)), avoiding duplication and computational efforts compared to traditional One-versus-One (OVO)/One-versus-All (OVA) methods. It works in association with all the proposed over_sampling techniques for handling multi-class datasets.

The notional flow chart of experimental execution is stated in fig.1. The steps involved in analysis framework are as:

1. Attaining a streaming input data (Apache Spark) using the Hadoop based mapreduce framework.
2. Building clusters (for assessing cluster cohesiveness) and a Random Forest tree of the data set.
3. Over_sampling the imbalanced data set to balance it.
4. Producing a model based on a new Random Forest tree and further analyze it.
5. Revising the model.
   - Using step 2 and 3, the newly updated data helps to improve Random Forest and can consequently be examined for cluster cohesiveness.
   - Repeat step 4 for real-time streaming input data set.

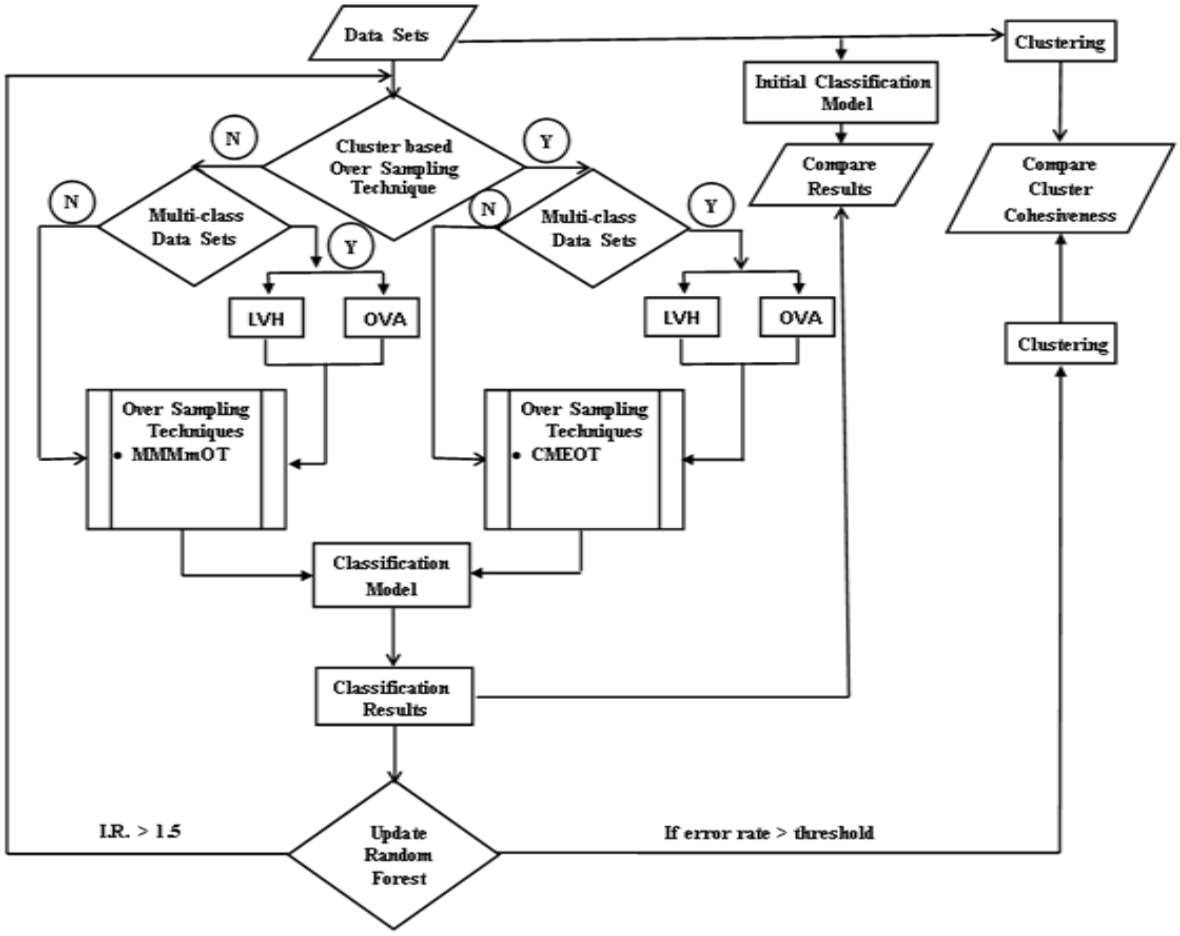

Figure 1. *Executional flow chart*

IV. PROJECTED TECHNIQUE: UPDATED CLASS PURITY MAXIMIZATION OVER_SAMPLING TECHNIQUE

a. UCPMOT:

The proposed technique processes under-over sampling on the clusters of individual classes [38- 39]. It helps to focus on low class purity clusters compared to their respective parents clusters, implicitly reducing the pure clusters beforehand. This technique improves the classification performance by tacitly addressing the 'between-class'/'with-in class' imbalances.

- Technique:

D_i – data set having 'N' instances
D_{mj} – majority class samples a_m (m = 1, 2, …., m)
D_{mn} – minority class samples b_n (n = 1, 2, …., n)
i – iteration count (1)
C_{ic} – intermediate clusters
D_o – a set of synthetic positive instances
BC- Binary class data set
M_n – Minority mediod instance from D_{mn}/respective clusters in-hand
M_j – Majority mediod instance from D_{mj}/respective clusters in-hand
D_{cp}– Degree of class purity

Compute safe levels of all samples [31] (based on no. of minority samples present in KNN of each individual instance).

Algorithm:

UCPMOT (D_i)

1. repeat
2. if $D_i = BC$
3. Select M_n and M_j
4. else
5. Select M_n (lowest minority class satisfying I.R.>1.5) and M_j (highest majority class)
6. end-if
7. Form clusters C_N and C_J around M_n and M_j respectively
8. if $i = 1$
9. if C_N or C_J != pure class
10. goto step 3 //for each impure cluster
11. else

12. goto step XX
13. else
14. if C_N or C_J != pure class
15. if $D_{cp}(C_N$ or $C_J) > D_{cp}$ (Parent)
16. goto step 3 //for each impure cluster
17. else
18. $C_{ic} = C_N$ and C_J
19. else
20. stop processing of C_N and C_J
21. end-if
22. append (D_o) = MMMmOT/CMEOT(C_{ic})
23. The classification is carried out on the final balanced data set

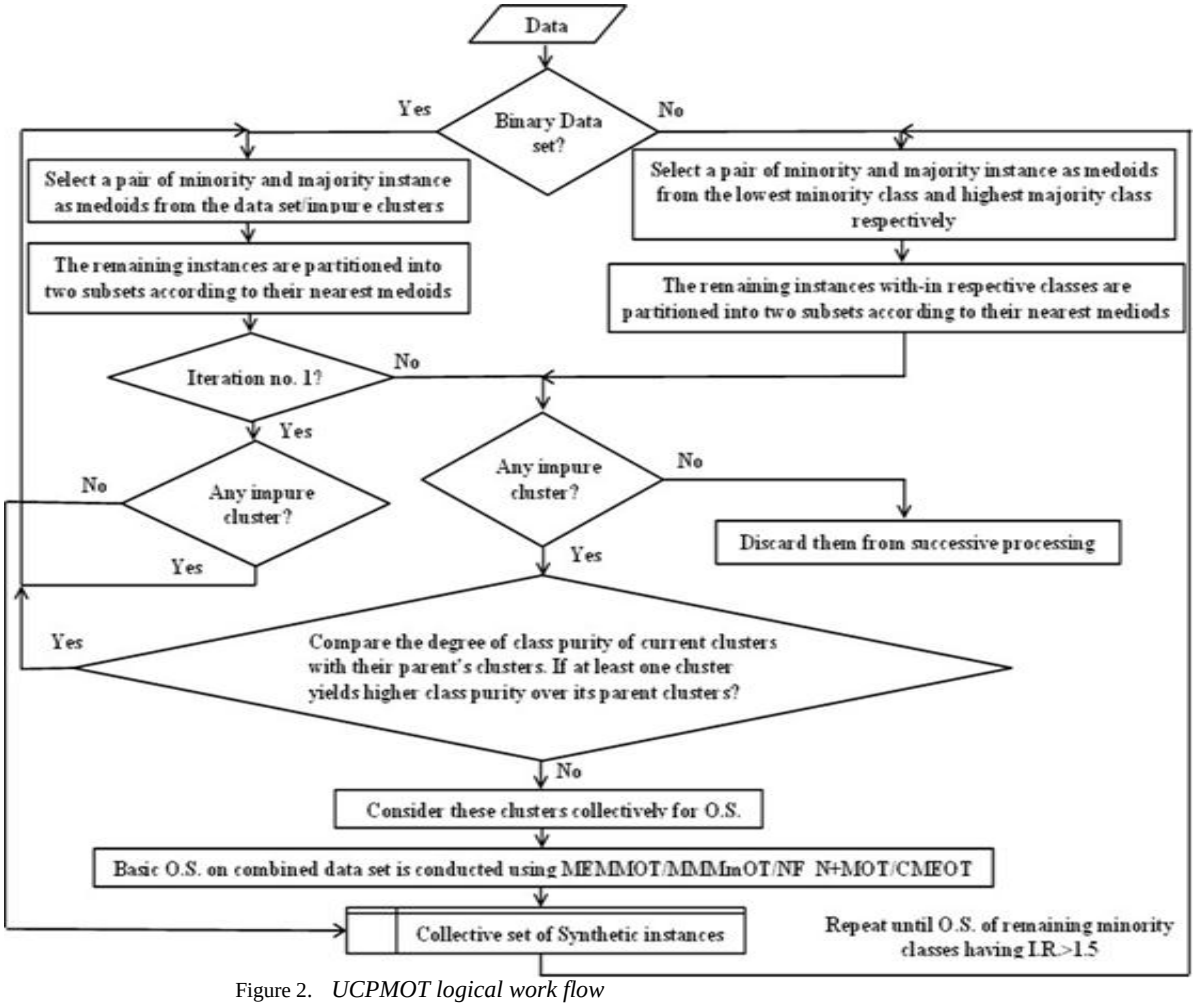

Figure 2. *UCPMOT logical work flow*

The executional drift of UCPMOT is depicted in fig. 2. It provides the logical trace of the over_sampling procedure on the imbalanced Big Data sets.

UCPMOT performs the fundamental over_sampling process in alignment with either of the two basic techniques specified beneath [30] (1 - non-cluster based and 1 - cluster based technique).

b. Non-cluster based technique: Minority Majority Mix mean Over_Sampling Technique (MMMmOT)

This technique is an exclusive progression of base technique (SMOTE). It considers both, minority and majority samples in KNN for further over_sampling. It relieves from low replicas and avoids the problem of overlapping samples.

D_{im} – intermediate synthetic samples
K_{NN} – 'K' number of nearest neigbhours
n – number of minority instances
S_Y – new synthetic sample
D.F. – displacement factor
SSS - Safe-Level based Synthetic Samples creation

Compute safe levels of all cases [31].
Algorithm (For 100% over_sampling rate):

1. **for** $D_{mn,}$ *i = 1 to n*
2. **for** *j =1 to* K_{NN}
3. $D_{mn}[i].KNN[j] =$ 'K_{NN}' *Nearest Neigbhour (KNN)*

4. **if** *KNN set = all minority instances*
5. **for** $m = 1$ *to* K_{NN} *and* $j = 1$ *to* K_{NN}
6. $D_{im}[m] = SSS[D_{mn}[i].KNN[j]$ *and* $D_{mn}[i]]$
7. S_Y = *average* ($D_{im}[m]$)
8. **if** S_Y = *duplicate*
9. *goto step 7 //delete the NN having a lowest safe level from the KNN including the interpolated instance from that instance*
10. $D_o = S_Y$
11. **if** *KNN set = all majority instances*
12. D_{im} = random (KNN set)
13. D_{imk} = minority NN (D_{im})
14. $S_{Y1} = SSS[D_{mn}$ *and* $D_{im}]$
15. $S_{Y2} = SSS[D_{mn}$ *and* $D_{imk}]$
16. S_Y = *average* (S_{Y1} *and* S_{Y2})
17. **if** S_Y = *duplicate*
18. *goto step 15 //search for the next NN from the data set*
19. $D_o = S_Y$
20. **else**
21. D_{im} = random (KNN set)
22. **if** D_{im} = minority instance
23. $S_Y = SSS[D_{mn}$ *and* $D_{im}]$
24. **if** S_Y = *duplicate*
25. *goto step 23 //search for the next NN from the KNN set or data set*
26. $D_o = S_Y$
27. **else**
28. D_{imk} = max. safelevel minority instance (KNN set or data set)
29. $S_{Y1} = SSS[D_{mn}$ *and* $D_{im}]$
30. $S_{Y2} = SSS[D_{mn}$ *and* $D_{imk}]$
31. S_Y = *average* (S_{Y1} *and* S_{Y2})
32. **if** S_Y = *duplicate*
33. *goto step 30 //search for the next minority NN from the KKN or data set*
34. $D_o = S_Y$

For over_sampling rate > 100%:
Repeatedly use the current over sampled set in-hand for over_sampling

OR

Choose (randomly or on safe level basis) an equal sample ratio from each over_sampling instance sets per iteration. Combine it with the base set of instances forming a new data set for the next over_sampling process

OR

Reiteration of step 2 to 4

For over_sampling rate < 100%:
Remove (randomly or considering highest safe levels) the interpolated samples satisfying the over_sampling rate

On failure, if any, in above cases regarding over_sampling rate, under-sampling based on clustering [32] can be planned to diminish majority classes. The over_sampling process is repeated for the remaining lowest minority classes belonging to multi-class data set satisfying I.R. > 1.5.

c. Safe-Level based Synthetic Samples creation (SSS)

The activity involved is as follows:

1. Find the safe level of minority instance under consideration & all its KNN instance.
2. Calculate the sum of safe levels of all instances in step 1.
3. Find the normalized value (N.V.) of safe levels for each individual instance in step 1.
   N.V. of instance = individual safe level value / total safe level value
   - N.V. is between 0-1
   - The summation of all individual N.V. equals to 1
4. The D.F. for synthetic sample creation in SMOTE processing:
   - If N.V. of the main instance under consideration is safer compared to N.V. of the chosen instance from KNN (randomly/S.L. based):
     D.F. = N.V. of chosen instance from KNN
   - If N.V. of the main instance under consideration is equal to N.V. of the chosen instance from KNN (randomly/S.L. based):
     D.F. = = 0.5
   - Else:
     i. If N.V. of chosen instance from KNN (randomly/S.L. based) < 0.5, then
     D.F. = 1 - N.V. of chosen instance from KNN
     ii. Else
     D.F. = N.V. of chosen instance from KNN

The proposed technique (SSS) will help to sensibly over sample the synthetic instances in a safe location. It will help to overcome the issues of overlapping and noisy regions in imbalanced data sets.

d. Cluster based technique: Clustering Minority Examples Over_Sampling Technique (CMEOT)

The technique involves only minority instances and is a wholesome cluster based technique. The computed cluster means are considered as new synthetic instances. It helps to address the data features of small disjuncts and lack of density. Moreover it compiles the objectives [33] of elevating centroids based over_sampling.
Compute safe levels of all cases [31].
Algorithm (For 100% over_sampling rate):

1. **for** $i = 1$ *to* c *// For c number of minority classes*

2. *$C_m[c]$ = clustering of each minority classes //K_{NN}<c using any clustering algorithm*
3. *new synthetic instances 'S_{Yset}' = computed medians of C_m*
4. **if** *S_{Yset}= duplicate*
5. *delete the respective instances*

For attaining over_sampling rate:

a. Repeat step 1 to 6 by adding the obtained medoids in-hand to current minority set

OR

b. Deletion of lowest safelevel minority instance (maintaining original data sets numbers) and reiterate step 1 to 6 (size_of_data set > K_{NN} and change in initial seeds)

The over_sampling process is repeated for the remaining lowest minority classes belonging to multi-class data set satisfying I.R. > 1.5.

V. EXPERIMENTAL SETTINGS

The objective of the experimental work is to authenticate the efficacy of proposed techniques. They are examined across three benchmarking techniques.

a. Details of data sets

The data sets under consideration are grouped into three categories viz. binary-class structured, multi-class structured and multi-class semi_unstructured data sets, each containing two data sets. They are from the standard UCI repository [37]. The details of data sets are given in Table I.

Table 1. Details of Data Sets

Category	Data set	#EX	#IR	#ATTR	#CL
Binary-class structured data sets	Skin	245057	3.81	4	2
	RLCP	5749132	273.67	12	2
Multi-class structured data sets	Car	1728	18.61	6	4
	KEGG-D	53413	13156.5	23	13
Multi-class semi-structured/ un-structured data sets	KDD Cup	4000000	3.99	42	24
	PAMAP2	3850505	14.35	54	19

b. Pre-settings and Assumptions

1. Enabling the 'noatime' option for mounting DFS.
2. Using a Lempel–Ziv–Oberhumer (LZO) compression techniques for intermediary data.
3. Allocating a suitable data type for the contents.
4. Converting the data sets contextually into numeric/symbolic structured forms.

c. Notations

The notations used in the experimental evaluation from Table III to VIII and Fig. 3 to 5 are noted in Table II.

Table 2. Notations

Notation	Algorithms	Notation	Data sets
A	SMOTE	**D1**	Skin
B	Safe-Level-SMOTE	**D2**	RLCP
C	ADASYN	**D3**	Car
D	UCPMOT_MMMmOT	**D4**	KEGG-D
E	UCPMOT_CMEOT	**D5**	KDD Cup
Notation	**Classifiers**	**D6**	PAMAP2
C1	Random Forest		
C2	Multilayer Perceptron		

VI. EXPERIMENTAL ASSESSMENT

The experimental evaluation is executed on four nodes Hadoop based mapreduce cluster. Each node has a configuration of Intel Core (TM) i7-4770 CPU@3.4 GHz with 8 GB RAM along with Ubuntu 14.04, Java 1.8.0 and Hadoop 2.7.4.

a. Comparison of F-Mesure and AUC values

The experiments are performed on six datasets [37] using LVH across two classifiers and keeping the value of cross-validation=10 and K=5. The results of the proposed technique (UCPMOT) are assessed using two parameters viz. F-measure and AUC values over three traditional techniques (SMOTE/Safe-Level-SMOTE/ADASYN).

Table 3. F-measure values (LVH)

Classifier	Dataset	Over_Sampling Techniques				
		A	***B***	***C***	***D***	***E***
C-1	D1	0.88	0.90	0.91	**0.94**	0.93
	D2	0.27	0.28	0.28	**0.74**	0.72
	D3	0.84	0.85	0.87	**0.92**	0.91
	D4	0.90	0.91	0.93	**0.96**	0.94
	D5	0.81	0.86	0.87	**0.91**	0.90
	D6	0.62	0.64	0.65	**0.77**	**0.75**
C-2	D1	0.80	0.82	0.84	**0.89**	0.88
	D2	0.25	0.26	0.26	**0.71**	0.70
	D3	0.83	0.85	0.86	**0.91**	0.90
	D4	0.88	0.90	0.91	**0.95**	0.93
	D5	0.79	0.84	0.85	**0.90**	0.89
	D6	0.60	0.62	0.63	**0.75**	**0.74**
Average		0.71	0.73	0.74	**0.86**	0.85

Table 4. AUC values (LVH)

Classifier	Dataset	Over_Sampling Techniques				
		A	***B***	***C***	***D***	***E***
C-1	D1	0.96	0.97	0.97	**0.99**	**0.99**
	D2	0.48	0.49	0.49	**0.81**	0.80
	D3	0.93	0.94	0.94	**0.96**	**0.96**
	D4	0.94	0.95	0.96	**0.99**	0.98
	D5	0.89	0.90	0.91	**0.95**	0.94
	D6	0.67	0.69	0.70	**0.83**	0.82
C-2	D1	0.91	0.92	0.93	**0.97**	0.96
	D2	0.47	0.47	0.47	**0.79**	**0.79**
	D3	0.90	0.91	0.91	**0.96**	0.94
	D4	0.93	0.94	0.95	**0.98**	0.97
	D5	0.87	0.89	0.90	**0.94**	0.93
	D6	0.65	0.67	0.68	**0.82**	0.81
Average		0.80	0.81	0.82	**0.92**	0.91

The average results of F-measure and AUC values, depict the superiority of UCPMOT over benchmarking techniques, representing improved classification. UCPMOT helps to address precise impure clusters in detail avoiding the trace of pure majority instances per cluster. UCPMOT_MMMmOT considers both category of instances while over_sampling to attain balanaced data set. UCPMOT_CMEOT realizes only the with-in clutser minority samples for over_sampling. Consideration of the heterogenous class structure leads UCPMOT_MMMmOT to achieve the highest results followed by UCPMOT_CMEOT. Additionally, the contextually structured data sets helps to notice encouraging results of C1 classifier compared to the C2. C2 miscarries the approximations of some linearly non-sperable minority instances.

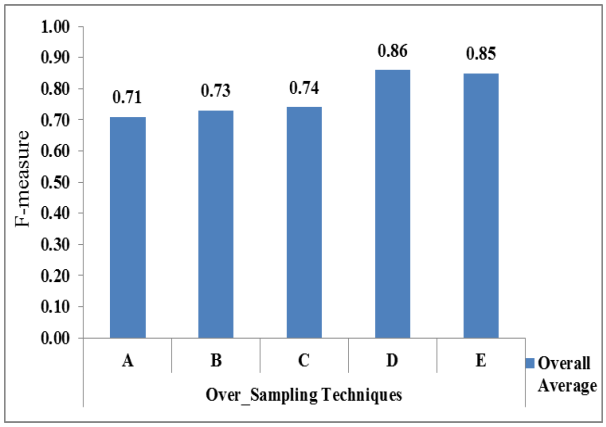

Figure 3. *Average F-measure values*

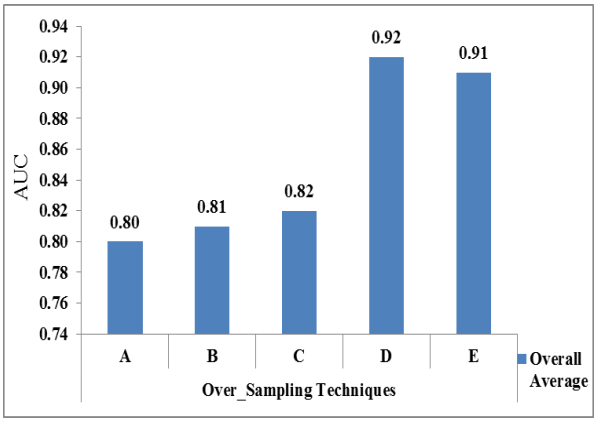

Figure 4. *Average AUC values*

The graph in fig. 3. and fig. 4., represent the average F-measure and AUC values respectively for all the techniques.

b. Comparison of LVH over OVA

Table V denotes the F-measure values for selected data sets (D1, D3 and D4) over OVA. The consequence of using LVH over OVA improving classification performance is specified in Table VI. to VIII. (Cross-validation=10 and K=5).

Table 5. F-measure values (OVA)

Classifier	Dataset	Over_Sampling Techniques				
		A	*B*	*C*	*D*	*E*
C-1	D1	0.88	0.91	0.92	**0.96**	0.94
	D3	0.85	0.85	0.88	**0.93**	0.92
	D4	0.91	0.92	0.95	**0.98**	0.96

LVH impressively handles the multi-class data sets for over_sampling. The consideration of highest majority class over all minority class reduces replication and avoids the overshooting issue. It implicitly overcomes the drawbacks of the OVA method for handling multi-class data sets. Table V shows marginal improvement of F-measure values compared to the results in Table III.

Table 6. Instance count (LVH)

Classifier	Dataset	Over_Sampling Techniques				
		A	*B*	*C*	*D*	*E*
C-1	D1	293675	286287	317854	279115	282698
	D3	4326	4107	4716	3154	4089
	D4	4762443	4581331	5101792	4313004	4423306

Table 7. Instance count (OVA)

Classifier	Dataset	Over_Sampling Techniques				
		A	*B*	*C*	*D*	*E*
C-1	D1	337889	323553	359985	311793	326880
	D3	4867	4653	5003	3472	4581
	D4	5892303	5613112	6114921	5119464	5543662

The Table VI and VII deliver the instance count of the data set after over_sampling using LVH and OVA respectively (F-measure).

Table 8. Analysis of LVH versus OVA (F-measure values)

Dataset	Over_Sampling Techniques									
	A		*B*		*C*		*D*		*E*	
	$\%RD_D$	$\%RD_F$	$\%RD_D$	$\%RD_F$	$\%RD_D$	$\%RD_F$	$\%RD_D$	$\%RD_F$	$\%RD_D$	$\%RD_F$
D1	14.0	0	12.22	1.1	12.43	1.09	11.06	2.1	14.49	1.06
D3	11.76	1.18	12.46	0	5.9	1.14	9.59	1.08	11.34	1.09
D4	21.20	1.1	20.24	1.09	18.06	2.12	17.09	2.06	22.48	2.1
Average	15.65	0.76	14.97	0.73	12.13	1.45	12.58	1.74	16.10	1.41

- $\%RD_D$: % relative difference in data set instances using OVA over LVH method in comparison to initial data set
- $\%RD_F$: % relative difference F-measure values (Random Forest – classifier) of OVA over LVH method in comparison to base F-measure values

The results in Table VIII authorizes the efficient implication of LVH method to handle the multi-class data sets. It analyzes the rise in % relative difference of over_sampling ratio over gain in performance. The relative regressional enhancement in F-measure values using OVA is less (average 1.5%) in relation to the progress of minority samples (13-15% compared to LVH).

The graph in fig. 5. positions the average values of % relative difference from Table VIII. X-axis represents the % relative difference (%RD_D versus %RD_F) and Y-axis represents the over_sampling techniques.

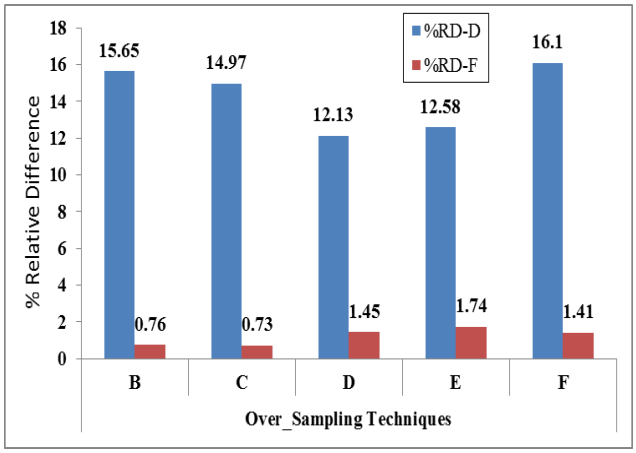

Figure 5. *Comparison of % RD_D versus %RD_F*

VII. DISCUSSION AND CONCLUSION

The paper compares various techniques for handling imbalanced Big Data sets. More explicitly, the enhanced clustered based technique UCPMOT in addition to SSS is proposed. The technique (non-cluster/cluster based) handles binary-class/multi-class data sets using LVH. It reduces bias and efficiently handles the issues related to several data characteristics like lack of density, small disjuncts and borderline instances. Experiments are carried out on standard data sets from UCI repository reveling wide-ranging of volume, attributes and I.R. The technique UCPMOT in combination with MMMmOT/CMEOT achieves improved F-measure and AUC values as stated in Table. 3 to 5 and 8. The results show an average 6-8% rise dictating the superiority of the proposed technique over benchmarking techniques. It helps to efficiently learn from imbalanced data sets. Two classifiers namely Random Forest and MultiLayer Perceptron are used for model building. The Random Forest classifier indicates a promising advancement in the results (2-3%) compared to MultiLayer Perceptron across all techniques (fig. 3. and 4.). Furthermore, the traditional data mining techniques are unable to survive with requirements urged by Big Data; hence, the Hadoop environment underlying the mapreduce framework is used to deal with it. The issues related to dataset shift and changing over_sampling rate needs to be further addressed in detail.

REFERENCES

[1] X. Wu et al., "*Data mining with big data*", IEEE Transaction on Knowledge and Data Engineering, Vol.26, Issue.1, pp.97–107, 2014.

[2] A. Gandomi, M. Haider, "*Beyond the hype: Big data concepts, methods, and analytics*" International Journal of Information Management, Vol.35, Issue.2, pp.137–144, 2015.

[3] D. Agrawal et al., "*Challenges and Opportunity with Big Data*", Community White Paper, pp.01-16, 2012.

[4] W. Zhao, H. Ma, Q. He., "*Parallel k-means clustering based on mapreduce*", CloudCom, pp.674-679, 2009.

[5] X.-W. Chen et al., "*Big data deep learning: Challenges and perspectives*", IEEE Access Practical Innovations: open solutions, Vol.2, pp.514 -525, 2014.

[6] "Big Data: Challenges and Opportunities, Infosys Labs Briefings - Infosys Labs," *http://www.infosys. com/infosys-labs/publications/ Documents/bigdata-challenges-opportunities.pdf.*

[7] N. Japkowicz, S. Stephen, "*The class imbalance problem: a systematic study*", ACM Intelligent Data Analysis Journal, Vol.6, Isuue.5, pp.429–449, 2002.

[8] H. He, E. Garcia, "*Learning from Imbalanced Data*", IEEE Transaction on Knowledge and Data Engineering, Vol.21, Isuue.9, pp.1263–1284, 2009.

[9] Y. Sun, A. Wong, M. Kamel, "*CLASSIFICATION OF IMBALANCED DATA: A REVIEW*", International Journal of Pattern Recognition Artificial Intelligence, Vol.23, Issue.4, pp.687–719, 2009.

[10] P. Byoung-Jun, S. Oh, W. Pedrycz, "*The design of polynomial function-based neural network predictors for detection of software defects*", Elsevier: Journal of Information Sciences, pp.40-57, 2013.

[11] V. López et al., "A*n insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics*", Elsevier: Journal of Information Sciences, Vol.250, pp.113–141, 2013.

[12] M. A. Nadaf, S. S. Patil, "*Performance Evaluation of Categorizing Technical Support Requests Using Advanced K-Means Algorithm*", IEEE International Advance Computing Conference, pp.409-414, 2015.

[13] R. C. Bhagat, S. S. Patil, "*Enhanced SMOTE algorithm for classification of imbalanced bigdata using Random Forest*" IEEE International Advance Computing Conference, pp.403-408, 2015.

[14] R. Sara, V. Lopez, J. Benitez, F. Herrera, "*On the use of MapReduce for imbalanced big data using Random Forest*", Elsevier: Journal of Information Sciences, pp.112-137, 2014.

[15] H. Jiang, Y. Chen, Z. Qiao, "*Scaling up MapReduce-based Big Data Processing on Multi-GPU systems*", SpingerLink Cluster Computing, Vol.18, Issue. 1, pp.369–383, 2015.

[16] G. Batista, R. Prati, M. Monard, "*A study of the behaviour of several methods for balancing machine learning training data*", ACM SIGKDD Explorations Newsletter - Special issue on learning from imbalanced datasets, Vol.6, Issue. 1, pp.20–29, 2004.

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

[18] H. Han, W. Wang, B. Mao, "*Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning*", Proceedings of the 2005 International Conference on Intelligent Computing, Vol.3644 of Lecture Notes in Computer Science, pp.878–887, 2005.

[19] B. Chumphol, K. Sinapiromsaran, C. Lursinsap, "*Safe-level-smote: Safelevel- synthetic minority over-sampling technique for handling the class imbalanced problem*", AKDD Springer Berlin Heidelberg, pp.475-482, 2009.

[20] H. He et al., "*ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning*", IEEE International Joint Conference on Neural Networks, pp.1322-1328, 2008.

[21] S. Garcia et al., "*Evolutionary-based selection of generalized instances for imbalanced classification*", Elsevier: Journal of Knowledge-Based Systems, pp.3-12, 2012.

[22] H. Feng, L. Hang, "*A Novel Boundary Oversampling Algorithm Based on Neighborhood Rough Set Model: NRSBoundary-SMOTE*", Hindawi: Mathematical Problems in Engineering, 2013.

[23] N. Chawla, L. Aleksandar, L. Hall, K. Bowyer, "*SMOTEBoost: Improving prediction of the minority class in boosting*", PKDD Springer Berlin Heidelberg, pp.107-119, 2003.

[24] H. Xiong, Y. Yang, S. Zhao, "*Local clustering ensemble learning method based on improved AdaBoost for rare class analysis*", Journal of Computational Information Systems, Vol.8, Issue.4, pp.1783-1790, 2012.

[25] F. Alberto, M. Jesus, F. Herrera, "*Multi-class imbalanced data-sets with linguistic fuzzy rule based classification systems based on pairwise learning*", Springer IPMU, pp.89–98, 2010.

[26] J. Hanl, Y. Liul, X. Sunl, "*A Scalable Random Forest Algorithm Based on MapReduce*", IEEE, pp.849-852, 2013.

[27] J. Kwak, T. Lee, C. Kim, "*An Incremental Clustering-Based Fault Detection Algorithm for Class-Imbalanced Process Data*", IEEE Transactions on Semiconductor Manufacturing, Vol.28, Issue.3, pp.318-328, 2015.

[28] S. Kim, H. Kim, Y. Namkoong, "*Ordinal Classification of Imbalanced Data with Application in Emergency and Disaster Information Services*", IEEE Intelligent Systems, Vol.31, Issue.5, pp.50-56, 2016.

[29] M. Chandak, "*Role of big-data in classification and novel class detection in data streams*", Springer Journal of Big Data, pp.1-9, 2016.

[30] S. Patil, S. Sonavane, "*Enhanced Over_Sampling Techniques for Imbalanced Big Data Set Classification*", Data Science and Big Data: An Environment of Computational Intelligence: Studies in Big Data, Springer International Publishing AG, Vol.24, pp.49-81, 2017.

[31] W. A. Rivera, O. Asparouhov, "*Safe Level OUPS for Improving Target Concept Learning in Imbalanced Data Sets*", Proceedings of the IEEE Southeast Conference, pp.1-8, 2015.

[32] S. Yen, Y. Lee, "*Under-Sampling Approaches for Improving Prediction of the Minority Class in an Imbalanced Dataset*", ICIC 2006, LNCIS 344, pp.731 – 740, 2006.

[33] C. Bunkhumpornpat, K. Sinapiromsaran, C. Lursinsap, "*DBSMOTE: Density-Based Synthetic Minority Over-sampling TEchnique*", Springer Journal of Applied Intelligence, pp.664-684, 2012.

[34] H. Guo et al.,"*Learning from class-imbalanced data: Review of methods and applications*", Elsevier Expert Systems With Applications, Vol.73, pp.220 – 239, 2017.

[35] Z. Zhang et al.,"*Empowering one-vs-one decomposition with ensemble learning for multi-class imbalanced data*", Elsevier Knowledge-Based Systems, Vol.106, pp.251 – 263, 2016.

[36] A. Vorobeva, "*Examining the Performance of Classification Algorithms for Imbalanced Data Sets in Web Author Identification*" Proceeding of the 18th Conference of FRUCT-ISPIT Association, pp.385 – 390, 2016.

[37] Machine Learning Repository, Center for Machine Learning and Intelligent Systems, US (NFS). *https://archive.ics.uci.edu/ml/datasets.html*

[38] K. Yoon, S. Kwek, "*An Unsupervised Learning Approach to Resolving the Data Imbalanced Issue in Supervised Learning Problems in Functional Genomics*", IEEE: International Conference on Hybrid Intelligent Systems, pp.1-6, 2005.

[39] M. Bach et al., "*The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis*", Elsevier Journal of Information Sciences, Vol.384, pp.174–190, 2017.

[40] D. Li et al., "*Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge*", Elsevier: Journal of Computation and Operational Research,Vol.34, pp.966–982, 2007.

[41] S. Barua et al., "*MWMOTE—Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning*", IEEE: Transaction on Knowledge and Data Engineering, Vol.26, pp.405–425, 2014.

[42] X. Ai et al., "*Immune Centroids Over-Sampling Method for Multi-class Classification*", T. Cao, E. Lim, Z. Zhou., T. Ho, D. Cheung, H. Motoda, Advances in Knowledge Discovery and Data Mining (eds), PAKDD 2015, Springer,Vol.9077, pp.251–263, 2015.

Authors Profile

Mr S S Patil received the B.E. degree in computer science and engineering and M. Tech. in computer science and technology from the Shivaji University, Kolhapur in 2003 and 2011 respectively. He is pursuing a Ph.D. degree in computer science and engineering under A.I.C.T.E. Q.I.P. scheme at Walchand College of Engineering (Govt. aided and an Autonomous Institute) affiliated to Shivaji University, Kolhapur, MH India.

Since 2010, he has been an Assistant Professor in the Computer Science and Engineering Department, Rajarambapu Institute of Technology, Rajaramnagar, MH – India. He has worked as head of Computer Science and Engineering department at Rajarambapu Institute of Technology, Rajaramnagar, MH – India. He is the author of a book chapter at Springer-Verlag and has more than 15 research papers. His research interests include Database Engineering and Big Data analytics. He has received a "Distinguished Facilitator" award at Inspire faculty contest organized by Infosys, Pune. He is a member of the IEEE.

Mrs S P Sonavane received her Bachelor Engineering (B.E.) and Masters (M.E.) degree in Computer Science and Engineering from Shivaji University, Kolhapur in 1992 and 2001 respectively. She has received a Ph.D. degree in 2010 in Computer Science and Engineering at Walchand College of Engineering (Govt. Aided_Autonomous Institute) affiliated to Shivaji University, Kolhapur, MH India.With two years of industry experience, she opted Teaching as a career profession.

Her Ph.D. work is supported under Young Scientist, research scheme by Department of Science and Technology, New Delhi, India. Dr. Shefali received Best Teacher Award in 2008 and is a member of many professional organisations.

Currently, she is working as an Associate Professor and heading Department of Information Technology at Walchand College of Engineering Sangli. She has received research funds from DST and AICTE for various technical projects promoting work in the area of Computer Vision and Information Security. She has a good number of publications in journals and participation in conferences with few IPR credentials at her account. She has extended her research interest further in the fields of Machine Learning and Big Data. She is an active member towards the implementation of Outcome Based Education (OBE) in engineering with special efforts in revamping the teaching methodology and its assessment.

