Cluster based Under_Over Sampling Technique to Handle Imbalanced Big Data Set Classification

Sachin Subhash Patil, Shefali Pratap Sonavane
Walchand College of Engineering, Sangli

ABSTRACT-The innovative opportunities of NoSQL Big Data have declared a new trial of frameworks. They assist in determining huge unidentified values from massive data sets. The cutting edge technologies with huge computing resources need a new pathway for enormous data generation, storage and processing. The volume and category of data created and warehoused are unimaginable and continues to grow. In addition, numerous applications treating imbalanced data sets have posed a priority of concern. Standard classifiers are not able to address the classification of over sampled imbalanced data sets using traditional techniques. An over_sampling technique: Majority Minority Cluster Based Under_Over Sampling Technique is suggested to improve classification and handling binary/ nonbinary-class imbalanced data sets. The nonbinary-class data sets are addressed using newly involved Lowest versus Highest method overcoming the challenges laid by traditional methods. The proposed technique is implemented with mapreduce environment on Apache Hadoop encompassing various data sets from the UCI repository. The technique constructing balanced steady data is next explored for classification and is authorized using parameters like G-mean and AUC. The experimental resultsachievedclearlymark the superiority of the presentedtechnique over the traditional techniques.

KEYWORDS-*Big Data, imbalanced data sets, Lowest versus Highest, multi-class, over_sampling techniques*

1. INTRODUCTION TO CLASSIFICATION OF IMBALANCED DATA SETS

ATA deficiency has outdrawn from the emerging huge digital world. Zettabytes of data is churning in and out per year. This gigantic varied data in forms of Volume, Velocity and Variety has led to a today's catchword 'Big Data'. To assimilate, exploit and further analyze this data has drawn research attention. Further the revision of performance practice is required to competently manipulate the conduct of streaming data.

Big Data challenges provoked by conventional data analytics are to handle competently. The superior verdict prediction from the inferred information out of massive, diverse data is a challenge [1]. Dealing effective economics and isolation of data is to be premeditated. Resources affirm [2], [3] the mass of digital data would be crossing Zettabytes by the year 2020 which is estimated to be 20 times more information than the current date. The crucial inclination of usage, mobility and deployment in addition to ecosystem capabilities has evolved down the line for Big Data management [3], [4], [5], [6].

Moreover in certain everyday applications, a lesser number of samples in one class compared to other classes has an escort to a condition called as a class disparity issue [7], [8], [9], [37]. Numerous real-world tribulations such as web author identification [39], medical judgment, scam recognition, finances, threat supervision, network invasion, software defect detection [10] have diverted attention towards analysis of concerns in imbalance nonbinary-class data sets. In a study of machine learning exploration, classifying correctly the negligible samples of such minority classes has become the main focus of study [11]. Traditional classifiers fail to predict precise classification for minority instances in imbalanced data sets ignoring their laxity in forming rule sets. The comprehensive pursuit to consider the consistency between the statistics of samples in each class leads to the challenges of learning from imbalanced data sets. The learning algorithms try to discover the preeminent result boundaries which are difficult to represent in imbalance data sets. Data characterized by skewed division is also an intact issue by diverse classifier learning algorithms.

In this paper, a better over_sampling (O.S.) technique viz. Majority Minority Cluster Based Under_Over Sampling Technique (MMCBUOST) dealing with imbalanced data of binary/multi-class problem is presented. The O.S. is carried out using two diverse techniques (Non-clustered based O.S. techniques) for improving classification. Further the classifiers viz. Naïve Bayes, AdaBoost and Random Forest (R.F.) [12], [13] are used to perform classification assessing their preciseness. The experiments are performed using the mapreduce based skeleton [14], [15]. The worthiness of techniques can fundamentally be evaluated using two measures: G-mean and AUC.

2. RELATED WORK

Classification of imbalanced data problem is addressed 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]. At data level, the focus is based on altering the volume of the original set for further analysis. The procedure level techniques work to revise a prevailing algorithm to promote processes dealing with imbalanced data. A mixer of data level and procedure level is integrated into a cost–sensitive technique to attain accuracy reducing the misclassification costs. The techniques

discussed in this paper deal with the data level.

Data level techniques are further distributed into three assemblies: Undersampling, O.S. and Hybrid technique [11], [14]. Every other technique does have its own advantages and disadvantages as like O.S. may tend to replicate noisy data or Undersampling might lose the significant data at insight. Random approach for both O.S. and undersampling is the simplest way to deal with the imbalanced data sets problem [16]. Correspondingly the O.S. results based on random approach emphasize the dominance over undersampling techniques. The proposed and allied techniques in this paper do basically work on O.S. style.

Synthetic Minority Oversampling Technique (SMOTE) algorithm [17] is one of the basic initial O.S. technique to deal with the imbalanced data set problem by synthesizing the marginal class examples. It aids to accomplish the required balance form. 'K' Nearest Neighbors (KNN) are selected on a random basis to satisfy the O.S. rate.

A SMOTE encounters various drawbacks, especially over-generalization, lack of addressing disjuncts, consideration of only minority class and applicability to binary-class. To overcome these drawbacks, techniques such as Borderline-SMOTE [18], Safe-Level-SMOTE [19] and Adaptive Synthetic Sampling (ADASYN) [20] were evolved. A Borderline-SMOTE helps to oversample only the minority examples near the borderline. Further, Safe-Level-SMOTE sensibly over samples minority instances nearby larger safe level improving classification accuracy. ADASYN progresses to learn and analyze data distribution by dropping the partiality. It adaptively changes the classification border near the hard examples to diagnose.

Evolutionary algorithms use the method belonging to nested generalized model considering objects in Euclidean n-space resolving the imbalance data set problem [21]. Neighborhood Rough Set Model based, SMOTE+GLMBoost [22] and NRBoundary-SMOTE are engaged in trading with boundary based oversampling. The ensemble methods viz. SMOTEBoost [23], AdaBoost [24] and RUSBoost are tangled with SMOTE to work for the problem of imbalanced data sets. Almost discussed techniques focus on the binary-class problem. F. Alberto, M. Jesus, and F. Herrera [25], proposed a fuzzy rule classification as a solution for the multi-class dilemma by merging the pairwise learning with preprocessing. The organization of ensemble based decision trees (R.F.) helps to effectively address classification algorithms [26], [38]. It comprises the attributes of scalability, durability and capable to handle continuous cum categorical data. J. Kwak, T. Lee and C. Kim [27], studied an incremental clustering based fault detection technique that comprises extreme class distributions of Gaussian/non-Gaussian types and process drifts. Ordinal classification of imbalanced data sets problem is the focus of discussion in [28], which approximates the class probability distribution using the weighted KNN method. Competent string based procedure to detect class in data streams is reflected in [29], including attributes of infinite-length and concept-evolution cum drift.

3. METHODOLGY

3.1 Architecture

The work involves the trial of functioning with imbalanced Big Data sets (I.B.D.) using statistical techniques. It leads to acquire data and manipulate it to a suitable format for further needful analysis of final results. The motto is to deliver a proportionate data satisfying the classification of imbalanced data sets.

The overall architecture illustrates the distinctive stages as shown in the Fig. 1. to investigate I.B.D [30]. It records and stores the varied streaming inputs and delivers some valuable primary comprehensions. Further the data is processed with O.S. techniques (Non-clustered/Clustered based techniques) to generate balanced data set for required analysis.

The clustered techniques need inputs for the number of clusters to be formed and its type. Prerequisite O.S. rate is provided to maintain imbalance ratio (I.R.) along with the pre-conditional value of 'k' for finding nearest neighbors. The distributed architecture is based on mapreduce framework capable of addressing heterogeneity and runtime scaling.

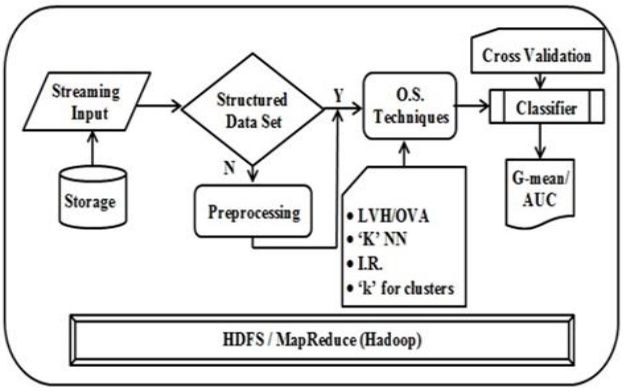

Fig. 1. Overall system architecture

3.2 LVH : Addressing Multi-class Imbalanced Data Sets

Traditional methods viz. One-versus-One (OVO) and One-versus-All(OVA) incur numerous overheads for handling multi-class imbalanced data sets. The conceived disadvantages of these two methods are as:

OVA –

1. Increased computation cost
2. Unable to accomplish O.S. for realistic need of classification satisfying all classes
3. O.S. may overshoot majority sub-classes
4. Leads to excess O.S. resulting in poor classification outcomes

OVO –

1. Converts a multi-class problem into binarization models inheriting its drawbacks

2. Hefty computation overhead

3. Borderline achievements toward improved accuracy

4. Discrete class balancing adds unwanted synthetic samples

5. One to one comparison increases model building time leading to affect final classification output

The current suggested method viz. LVH suffices most of the disadvantages of OVO and OVA method. This method benefits to improve classification performance including concentrated computation.

LVH [30]- The hint cited for this method is more forceful for treating multi-class domains. The advantages are as:

1. O.S. considers a single highest majority class to be compared individual minority class satisfying a certain threshold (I.R. >1.5)

2. Reduces computation and O.S.

3. Accurate synthetic O.S. avoiding duplication cum overrunning any majority classes

4. Fulfills implicitly the O.S. need of residual majority classes conforming the target majority class

LVH is well illustrated with an example. A data set is considered having five classes P, Q, a, b and c. P and Q are assumed to be majority classes and remaining others as minority classes. Class P is supposed to be the highest majority class and 'a' as the lowest minority class. LVH first of all, leads to balance the lowest minority class (a). It deals with only the highest majority class (P) complying the O.S. rate for balancing the same. Likewise, the left over minority classes (b and c satisfying I.R.>1.5) are further Over_sampled individually with respect to 'P' to finally balance the data set.

The LVH method is used to handle multi-class data sets with all the newly proposed O.S. techniques discussed in this paper.

3.3 MMCBUOST: O.S. Technique for Improving Classification Outcomes

The design of this technique is to reflect on intra and inter-class imbalances consecutively. It performs under-over sampling beforehand on the group of individual classes. This technique categorizes under clustered based technique.

Technique –

Let the data set be D_i having 'N' instances, D_{mj} – majority class samples a_m (m = 1,2,….,m) and D_{mn} – minority class samples b_n (n = 1,2,….n).

Compute safe levels of all samples [31].

Algorithm:

Input: a set of all instances D_i

Output: a set of all synthetic positive instances D_o

1. $D_o = \emptyset$

2. Repeat {

3. Check D_i is binary-class data set:

4. if Yes

5. Form the clusters of D_{mj} and D_{mn} as C_{mj} and C_{mi} respectively (e.g. using K-means and assuming a number of clusters of D_{mj}> number of clusters of D_{mn} +2).

6. Find the immediate large majority class cluster (Mj_l) compared to the highest minority class cluster (m_{nl}).

7. else

8. Form the clusters of highest majority class (C_{hmj}) and all individual minority classes (below 40%) from the dataset (e.g. using K-means and assuming a number of clusters of D_{mj}> number of clusters of D_{mn} +2).

9. Select the most minority class clusters (C_{mmi}). Find the immediate large majority class cluster (Mj_l) from the C_{hmj}, compared to the highest minority class (m_{nl}) under consideration.

10. Undersample all the majority clusters (of C_{hmj} only) above $Mj_l$to the level of Mj_l (based on safe level or any other technique for e.g. SBC [21]). [Assuming the number of minority instances almost per cluster does not meet the m_{nl} under consideration]

11. Calculate the complete number of majority class instances 'v' after undersampling.

12. Calculate the required O.S. rate 'o' centered on 'v' (where I.R.<=1.5).

13. Compute the distinct minority cluster O.S. rate under consideration based on 'o' complying I.R. (Assuming to equalize the number of each cluster sample instances).

14. Basic O.S. on distinct minority clusters is conducted either in association with MEre Mean Minority Over_Sampling Technique (MEMMOT)/ Adjacent Extreme Mix Neigbours Over_Sampling Technique (AEMNOST) (Majority sample from inclusive set of majority class formed in step 10. and minority samples within the same cluster are to be considered for the finding of nearest neighbor while O.S. process).

15. Add the synthetic instances to original minority class under consideration (C_{mi}/C_{mmi}) and D_o.

16. } Until O.S. of remaining minority classes as per step 8. and 9.

17. return D_o

18. The classification is further carried out on the final balanced data set.

The stated technique (MMCBUOST) works in alignment to either of the two basic techniques (MEMMOT/AEMNOST) for elementary O.S. process [30]. They help in context to address binary/multi-class imbalanced data sets and are basically categorized into non-cluster based techniques.

Technique-1 - MEMMOT

Compute safe levels of all cases [31]. For every minority instance b_n under consideration.

Algorithm (For 100% O.S. rate):

1. Search KNN for all instances.

2. Compute SMOTE individually with all the KNN instances.

3. New synthetic instance 'S_Y'= average (all interpolated instances).

4. S_Y = duplicate instance? Yes, delete the NN having a lowest safe level from the KNN including the interpolated instance from that instance. Go to step 2.

a. For O.S. rate > 100%:

Repeatedly use the current over sampled set in-hand for further O.S.

OR

Randomly/Considering safe levels choose an equal sample ratio from each O.S.instance sets per iteration. Combine it with the base set of instances forming a new data set for next O.S. process.

OR

Reiteration of step 2 to 4.

b. For O.S. rate < 100%:

Randomly/Considering lowest safe levels remove the interpolated samples satisfying the O.S. rate.

In view of the failure to above cases, under-sampling based on clustering [32] can be planned to diminish majority classes.

This technique delivers improved classification with comprehensive interpolated minority instances.

Technique-2 - AEMNOST

Technique intends to study the effect of an equalmixture of nearest, farthest and the middle element for forming synthetic instances. It may help to provide a wide range of inputs avoiding overlapping and replication along with improving classification.

Compute safe levels of all cases [31]. For every minority instance b_n under consideration.

Algorithm (For 100% O.S. rate):

1. Search K instances such that

a. K/2 nearest

b. K/2 farthest and

c. midpoint element (except for the even value of 'K')

where N>1 and k <= N

2. Search KNN for all instances.

3. If all KNN instances are:

a. minority – follow step 4 and 5.

b. majority – follow step 6 and 7.

c. Else – follow step 8 and 9.

4. Compute SMOTE individually with all the KNN instances. New synthetic instance 'S_Y'= average (all interpolated instances).

5. S_Y = duplicate instance? Yes, delete the NN having a lowest safe level from the KNN including the interpolated instance from that instance. Go to step 4.

6. Select any random instance from KNN and search for its nearest minority instance. Compute interpolated instance from both independently with the main instance under consideration using SMOTE. New synthetic instance 'S_Y'= average (both interpolated instances).

7. S_Y = duplicate instance? Yes, search for its next nearest minority instance. Go to step 6.

8. Select any random instance from KNN.

a. If minority - Compute SMOTE.

b. If majority - Search for a maximum safe level minority instance within the KNN set. Compute interpolated instance from both independently with the main instance under consideration using SMOTE. New synthetic instance 'S_Y'= average (both interpolated instances).

9. S_Y = duplicate instance? Yes, search for its next nearest minority instance. Go to step 8.

For required O.S. rate > or < 100%, the same strategy discussed in MEMMOT based O.S. technique is planned to be used.

4. EXPERIMENTAL CONTEXT

The objective of the research is to verify the effectiveness of the proposed techniques dealing with the problem of class imbalance in Big Data sets. The experimental setup and investigation are presented herewith comparing diverse techniques.

4.1 Details of Data Set

The five standard data sets from UCI repository [40] are selected for experimental analysis. The data sets considered are categorized into two groups. They characterize with a varied number of instances from lower to higher quantum, wide-ranging of attributes and comprehensive I.R. from low to high significance. The details of data sets are as follows:

Table 1.Features of Data set

Category	Data set	#EX	#IR	#ATTR	#CL
Binary-class data sets (B)	RLCP	57,49,132	273.67	12	2
	Skin	2,45,057	3.81	4	2
	Nomao	34,465	2.51	120	2
Multi-class data sets (M.C.)	KDD	40,00,000	3.99	42	24
	PAMAP	38,50,505	14.35	54	19

Note:

- #I.R. - Highest majority class w.r.t. lowest minority class (for multi-class data sets)
- A set of useful attributes is considered for experimentation

Table 1. summarizes the particulars of chosen data sets including its category, standard name, the number of instances (#EX), I.R. (#IR), the number of attributes (#ATTR) and the number of classes (#CL).

4.2 Pre-settings and Assumptions

1. RAID/LVM are avoided on TaskTracker/DataNode systems.

2. 'noatime' option is used for mounting DFS and MapReduce storage.

3. Using compression techniques (LZO) for intermediate data.

4. Almost data sets are contextually converted into numeric/symbolic structured standards for further study.

5. The number of mapper tasks is maintained in ratio to some multiple of mapper slots in the cluster to effectively utilize the slots.

4.3 Notations

The notations used henceforth for classifiers, data sets and algorithms during experimental analysis are given in Table2.:

Table 2.Notations

Classifiers	**D4** - KDD
C1 - Multilayer Perceptron(M.L.P.)	**D5** - PAMAP
C2 - AdaBoostM1 (Ad.B.)	*Algorithms*
C3 - Random Forest (R.F.)	**A** - Original data set result
Data sets	**B** - SMOTE
D1 - RLCP	**C** - Safe-Level SMOTE
D2 - Skin	**D** - MMCBUOST_MEMMOT
D3 - Nomao	**E** - MMCBUOST_AEMNOST

5 Experimental Analysis

The techniques are applied to two class/multi-class data sets. The evaluation of techniques to handle imbalanced classification is planned using two measures viz. G-mean and AUC.The experimented G-mean and AUC values are obtained on over_sampled data sets by performing 10-fold cross-validation using k=5. Three classifiers namely M.L.P., Ad.B.and R.F. are used in the experimental work. The outcomes are compared between benchmark (SMOTE/Safe-Level SMOTE) and proposed technique (MMCBUOST).

The trialing is conducted on 12 node Hadoop clusters with two master nodes (Namenode and Job tracker) and 10 slave nodes. Each node has an Intel Core (TM) i7-4770 CPU@3.4 GHz having 8 GB RAM. The cluster works on Ubuntu 14.04, Java 1.8.0 and Hadoop 2.6.4.

The results obtained demonstrate overall better average values of G-mean and AUC for the proposed technique MMCBUOST in combination with MEMMOTand AEMNOST representing enhanced classification. R.F. signifies encouraging results for almost all techniques compared to other two classifiers (M.L.P. and Ad.B.).

5.1 Binary-class Data Sets (B)

The results comprising G-mean values for the three data sets are presented in Table. 3 (CL: Classifier, DS: Data set, OA: Overall Average, MA: Multi-class Addressing method, MP: Number of mappers). The results of the plain original data set in comparison to benchmarking and proposed techniques are noted for observation. There is a marginal growth of classification improvement in benchmarking techniques. It is apparent from the outcomes that the projected technique MMCBUOST in combination with MEMMOTand AEMNOST represents improved classification results.

Table 3.G-Mean Values for Binary-class data set

CL	DS	O.S. Techniques				
		A	**B**	**C**	**D**	**E**
C1	D1	0.21	0.23	0.24	0.63	0.59
	D2	0.80	0.81	0.83	0.91	0.91
	D3	0.82	0.84	0.86	0.92	0.90
C2	D1	0.21	0.24	0.25	0.67	0.61
	D2	0.84	0.88	0.89	0.95	0.93
	D3	0.83	0.85	0.87	0.93	0.93
C3	D1	0.23	0.25	0.25	0.68	0.62
	D2	0.84	0.89	0.91	0.97	0.96
	D3	0.85	0.89	0.91	0.95	0.94
OA		0.63	0.65	0.67	**0.85**	**0.82**

The graph in Fig 2 illustrates the aggregated average of G-mean values for all techniques under consideration. The binary-class data sets are introspected using three classifiers viz. M.L.P., Ad.B. and R.F.

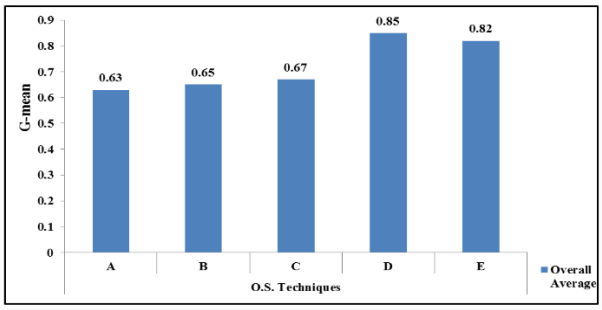

Fig 2: Average G-mean values for B

In Fig. 2.,x-axis represents the O.S. techniques (B-E) including bare results (A) and y-axis represents the values for G-mean. Analyzing the graphs, technique MMCBUOST+MEMMOT achieves the higher values of G-mean compared to all other techniques for almost all classifiers followed by MMCBUOST+AEMNOST.

5.2 Multi-class Data Sets (M.C.)

The Table 4.exhibits the performance of all O.S. techniques over LVH in terms of AUC values.

Table 4.AUC Values for Multi-class data sets

CL	DS	O.S. Techniques				
		A	B	C	D	E
C1	D4	0.69	0.87	0.89	0.90	0.91
	D5	0.42	0.61	0.66	0.72	0.69
C2	D4	0.70	0.88	0.91	0.92	0.92
	D5	0.43	0.63	0.67	0.73	0.72
C3	D4	0.72	0.89	0.90	0.92	0.93
	D5	0.44	0.63	0.68	0.75	0.74
OA		0.57	0.75	0.79	**0.82**	**0.82**

MMCBUOSTachieves higher results of AUC values in Table 4. compared to all other techniques. The results clearly demonstrate the significance of LVH for handling I.B.D.

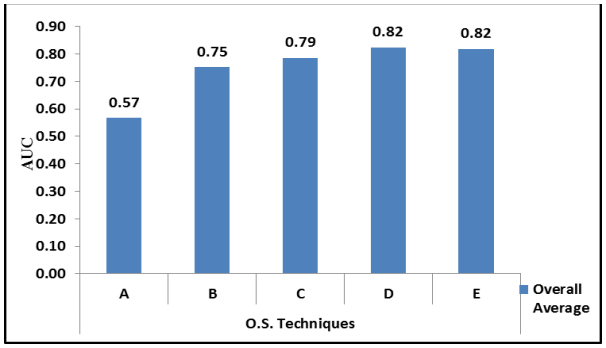

Fig3: Average AUC values for M.C. (LVH)

The Fig3 describes the overall average values of AUC enclosing LVH. The x-axis represents the O.S. techniques and the y-axis represents the average AUC values using three classifiers. SMOTE and SafeLevel SMOTE techniques show poor performance in comparison to proposed new technique.

6. Conclusion

In this paper, the several techniques for handling I.B.D. are compared. More explicitly, the attempt is made to propose an advanced clustered based technique MMCBUOST in addition to LVH method which is able to deal with binary-class/multi-class Big Data. It helps to reduce bias and efficiently handle the drawbacks of traditional techniques in alignment to improve classification results. The issues upraised due to fundamental data characteristics like overlapping cum influence of borderline instances, lack of density and small disjuncts are addressed effectively. The M.L.P. and well-known decision tree ensemble classifiers are used for model building and analysis. Hadoop environment underlying mapreduce framework is used to treat the necessities pressed by Big Data management. Experiments are carried out on standard data sets from UCI repository. Data sets under consideration exhibit wide-ranging of I.R., data size and a number of attributes; thus catering a diverse test bed. The MMCBUOST combines the power of MEMMOT and AEMNOST to massively improve precision and recall implicitly achieving a better G-mean and AUC values. The experimental results provide the validation that the proposed technique can efficaciously be used for learning from I.B.D.

REFERENCES

[1] X. Wu et al., "Data mining with big data," *IEEE Trans.on Knowledge and Data Engg.*, vol. 26, no. 1, pp. 97–107, 2014.
[2] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, no. 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, and 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 Intelli. Data Analysis Journal*, Vol. 6, no. 5, pp. 429–449, 2002.
[8] H. He, E. Garcia, "Learning from Imbalanced Data," *IEEE Trans. on Knowl.and Data Engg.*, Vol. 21, no. 9, pp. 1263–1284, 2009.
[9] Y. Sun, A. Wong, M. Kamel, "CLASSIFICATION OF IMBALANCED DATA: A REVIEW," *Int. Journal of Pattern Recognition Artificial Intelligence*, Vol. 23, no. 4, pp. 687–719, 2009.
[10] P. Byoung-Jun, S. Oh, and 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., "An 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, and 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, and Z. Qiao, "Scaling up MapReduce-based Big Data Processing on Multi-GPU systems," *SpingerLink Cluster Computing*, vol. 18, no. 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, no. 1, pp. 20–29, 2004.
[17] N. Chawla, K. W. Bowyer, L. O. Hall, and 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, and 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, and 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, and K. Bowyer, "SMOTEBoost: Improving prediction of the minority class in boosting," *PKDD Springer Berlin Heidelberg*, pp. 107-119, 2003.
[24] H. Xiang, Y. Yang, and S. Zhao, "Local clustering ensemble learning method based on improved AdaBoost for rare class analysis," *Journal of Computational Information Systems*, Vol. 8, no. 4, pp. 1783-1790, 2012.
[25] F. Alberto, M. Jesus, and 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, and 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, no. 3, pp. 318-328, 2015.
[28] S. Kim, H. Kim, and Y. Namkoong, "Ordinal Classification of Imbalanced Data with Application in Emergency and Disaster Information Services," *IEEE Intelligent Systems*, vol. 31, no. 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," in *Data Science and Big Data: An Environment of Computational Intelligence: Studies in Big Data, Springer International Publishing AG*, 2017, ch. 3, Vol. 24, pp. 49-81.
[31] W. A. Rivera, O. Asparouhov, "Safe Level OUPS for Improving Target Concept Learning in Imbalanced Data Sets," Proceedings of the IEEE Southeast Con., pp. 1-8, 2015.
[32] S. Yen and 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] *http://www.causality.inf.ethx.ch/data/SDIO.html*
[35] M. Weiss, S. Sari and N. Noori, "Niche formation in the mashup ecosystem," *Technology Innovation Management Review*, 2013.
[36] H. Rong, D. Wanchun and L. Jianxun, "ClubCF: A Clustering-Based Collaborative Filtering Approach for Big Data Application," *IEEE Transactions on Emerging Topics in Computing*, Vol. 2, no. 3, pp. 302 – 313, 2014.
[37] 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.
[38] Z. Zhang et al.,"Empowering one-vs-one decomposition with ensemble learning for multi-class imbalanced data," *Elsevier Knowle dge-Based Systems*, Vol. 106, pp. 251 – 263, *2016.*
[39] 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.
[40] *https://archive.ics.uci.edu/ml/datasets.html*