Applied Computer Systems

ISSN 2255-8691 (online)
ISSN 2255-8683 (print)
December 2019, vol. 24, no. 2, pp. 104–110
https://doi.org/10.2478/acss-2019-0013
https://content.sciendo.com

A New Big Data Model Using Distributed Cluster-Based Resampling for Class-Imbalance Problem

Duygu Sinanc Terzi[1], Seref Sagiroglu[2*]
[1, 2]*Department of Computer Engineering, Gazi University, Ankara, Turkey*

***Abstract* – The class imbalance problem, one of the common data irregularities, causes the development of under-represented models. To resolve this issue, the present study proposes a new cluster-based MapReduce design, entitled Distributed Cluster-based Resampling for Imbalanced Big Data (DIBID). The design aims at modifying the existing dataset to increase the classification success. Within the study, DIBID has been implemented on public datasets under two strategies. The first strategy has been designed to present the success of the model on data sets with different imbalanced ratios. The second strategy has been designed to compare the success of the model with other imbalanced big data solutions in the literature. According to the results, DIBID outperformed other imbalanced big data solutions in the literature and increased area under the curve values between 10 % and 24 % through the case study.**

***Keywords* – Big data, cluster-based resampling, imbalanced big data classification, imbalanced data.**

I. Introduction

Big data is composed of many miscellaneous and autonomous resources with various dimensions and complex relationships that are beyond the capacity of traditional techniques and tools [1]. Big data, which has become a more important production factor than material assets, has a great potential to create value and insight when the challenges are overcome. The challenges lie at different levels including: acquisition, storage, exploration, visualization, sharing, analysis, management, and security of data [2].

The imbalanced data, one of the common big data challenges, is caused by real world applications producing classes with different distributions. The first type of class that is under-presented with fewer instances than others because of the rare events, abnormal patterns, unusual behaviours, or interruptions during gathering of data is known as the minority, while the remaining class/classes that have an abundant number of instances are named as majority [3]. Figure 1 maps the types of imbalanced data [4], frequently suggested solutions in the literature [5], assessment metrics to evaluate effectiveness of these solutions [6], and widespread real-world applications of imbalance data [3].

Traditional data management methods work typically on the assumption of uniformly represented class distributions, equally expressed sub-concepts in classes, and correctly defined attributes and labels. Therefore, the final model is generally assumed to be accurate. However, practically imbalanced data overwhelms the learning processes of algorithms and creates bias towards majority class in accuracy, although minority class prediction is more important and costly. For instance, detecting an attack is more important than detecting normal traffic, or diagnosing the disease is more critical than diagnosing health. Class imbalanced problem is typically handled in three ways: under/oversampling, modifying algorithm, and reducing misclassification cost [5]. However, these approaches have several limitations, such as working well on small data, having more computing and storage costs because of algorithm complexity, being slow by algorithm's nature, handling either binary-class or multi-class problems, and requiring predefined threshold values. When these issues are considered again in the context of big data, mostly similar suggestions have been developed and transformed into MapReduce procedures for improving performance in high volume, diverse, and complex data.

The proposed approaches on imbalanced data classification in big data may be grouped as data-level, algorithm-level, and cost-sensitive solutions [7]. Data-level solutions usually consist of applying one or more base classification algorithms after rebalancing data. Algorithm-level solutions include enhancements for learning stage. Cost-sensitive solutions provide metrics and methods that are suitable for class distribution. Various data-level approaches can be summarised as traditional resampling techniques that are adapted [8] or enriched [9] in MapReduce workflow, combination of metric learning algorithms and balancing techniques [10], improving several big data supervised techniques [11], application of random over-sampling with evolutionary feature weighting [12], [13], evolutionary under-sampling methods embedded in MapReduce scheme [14], [15], a MapReduce-based data reduction scheme [16], a MapReduce design based on Neighborhood RoughSet Theory [17], lastly, an enhancement for multi-class imbalanced classification [18]. Several methodologies that include algorithmic modifications for cost-sensitive learning can be enumerated as maximization of gmean which results in a non-convex loss function [19], a cost-sensitive support vector machine [20], an instance weighted variant of support vector machine [21], a

[*] Corresponding author's e-mail: ss@gazi.edu.tr

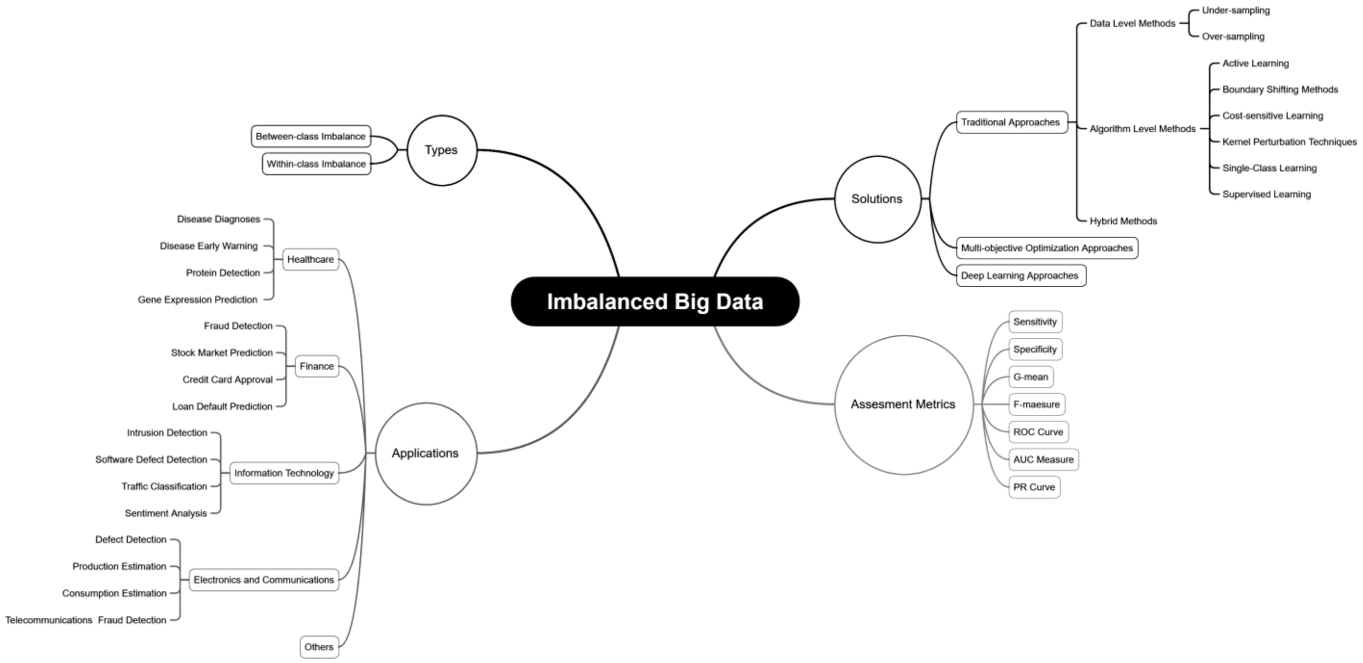

Fig. 1. Map of imbalanced big data.

MapReduce implementation of linguistic fuzzy rule-based classification [22], [23], lastly, extreme learning machine algorithms [24], [25]. In addition to these solutions, there are also frameworks using many approaches and algorithms in a holistic way [26]. When the studies in the literature are compared, each approach has superiority in terms of different aspects and in most cases there is an inevitable trade-off between the complexity of the analysis model and the difficulty in classifying the data. However, literature is inadequate regarding within-class imbalance or small disjuncts problem. The minority class has much lower rate in big data, which complicates the learning process, and the sub-concepts in the minority class cannot be expressed well by the developed model. Since most classifiers create large disjuncts and cause difficulties to detect sub-concepts, cluster based resampling methods gain importance in within-class imbalance [27].

Therefore, in the present paper, we focus on the class imbalanced problem and solve it with a novel resampling model called Distributed Cluster-based Resampling for Imbalanced Big Data (DIBID). The DIBID has been designed to effectively overcome both between-class and within-class imbalance problems in the big data, especially when faced with the challenge of volume.

The rest of the paper is organised as follows: The approaches used in the proposed MapReduce design are summarised in Section II. In Section III, DIBID is described in detail. The experimental studies and the obtained results are evaluated in Section IV. The proposed model is discussed in Section V. Lastly, the paper is concluded in Section VI.

II. Preliminaries

The effective pre-processing techniques enable better data utilisation and better models by eliminating the irregularities in big data. The class imbalance problem, one of the common data irregularities, causes the development of under-represented models. As imbalanced data hosts several behaviours and characteristics, the developed methods need to focus on the solution of the underlying causes that make the problem more difficult. For this purpose, DIBID is elementarily comprises of three methods: clustering, resampling, and classification.

It is difficult to decide which clustering algorithm is best suited for a particular big dataset, because of the difficulties to find out the benefit of one algorithm over another with respect to both theoretical and empirical perspectives. At this stage, many clustering approaches can be used. *k*-means clustering algorithm, which is easily applicable and effectively detects the condensed areas, is preferred in the present to detect small disjuncts that cause within-class imbalance. Various resampling techniques, such as RUS (Random Under-sampling), NearMiss, ROS (Random Over-Sampling), or SMOTE (Synthetic Minority Over-Sampling Technique) can be used on clustered data set parts and the performance of each varies according to data distribution. Due to the complexity of some resampling methods and the difficulty of adapting some techniques to big data analysis, RUS, ROS, and SMOTE are used in the proposed model. The classification, which aims at building a concise model of the distribution of predictor labels, requires continuous collection of data and learning the characteristics from big data. With the increase of

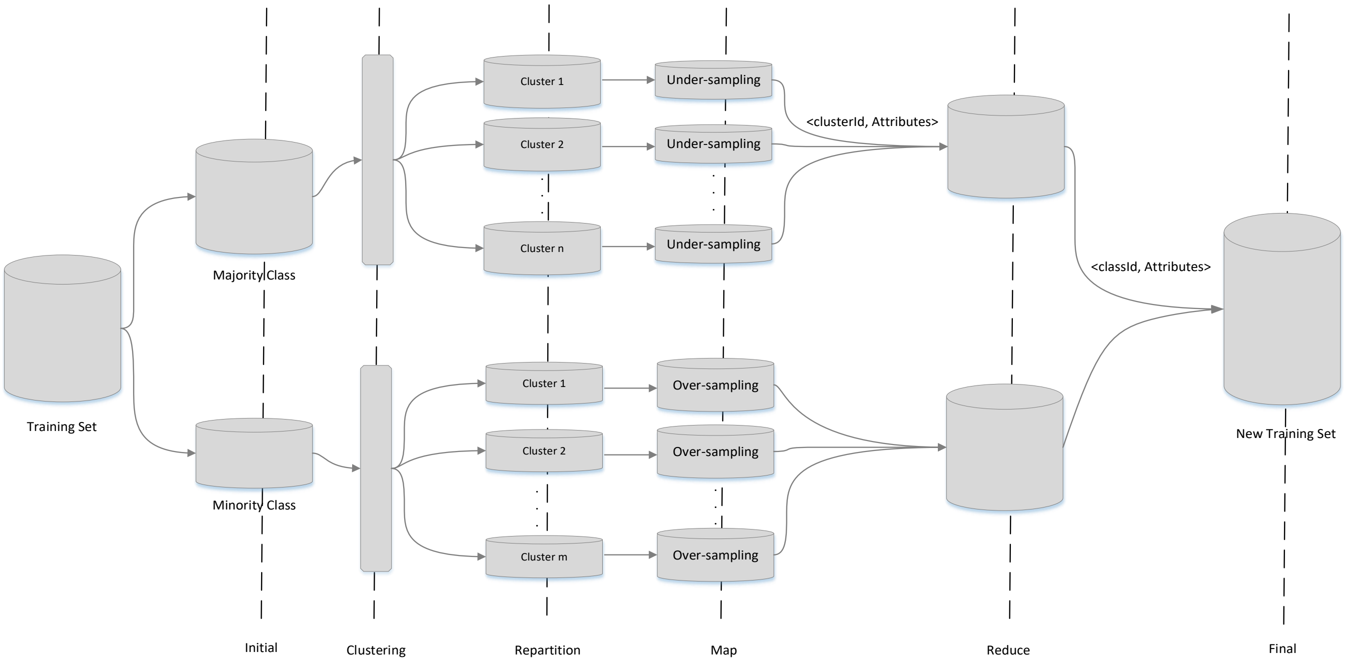

Fig. 2. Flowchart of the MapReduce design.

data, the model is expected to cover all characteristics, but this situation forces the capacity of most classifiers. Random Forest (RF) is a very popular decision tree ensemble that is used in classification due to its outstanding performance. Therefore, it is preferred in the present study.

In the selection of the evaluation metric of the model created after all analyses, accuracy, which assumes equal error costs and relatively balanced class priorities, cause worse predictions than receiver operating characteristics analysis-based assessments that make no assumptions about costs or priorities. Consequently, Area Under the Curve (AUC) is used to evaluate classification performances of DIBID.

III. Distributed Cluster-Based Resampling for Imbalanced Big Data (DIBID)

As mentioned before, the aim of the study is to produce a solution for both between-class and within-class imbalance classification problem on big data sets in a scalable way. Therefore, DIBID is proposed as the combination and improvement of such methods outlined below:

- the data set is divided into two as training set and test set before the resampling process. Although it may seem to reduce the representation in training data and undermine the classifier's predictions [28], it is aimed at preventing the leakage of information;
- minority and majority classes are clustered separately [29] to determine different behaviours in imbalanced big data;
- minority classes are over-sampled and majority classes are under-sampled to reduce the imbalance ratio between classes [30];
- cluster based under-sampling for majority class [31] and cluster based over-sampling for minority class [32] are implemented to procure quality of instances in small disjuncts;
- unlike the proposed methods, which generally increase a fewer number of classes up to a maximum number of classes, in this model, over-sampling and under-sampling ratios are used to prevent overfitting and to avoid throwing away useful information [33];
- resampling methods are executed on partitions to parallelise distributed data processing for improving the performance and reducing the network traffic [34].

The presence of sub-behaviours in classes is implicit in most cases and it increases the learning complexity. Thanks to DIBID, resampling is performed in accordance with the sub-behaviours in classes, while the problem is transformed into sub-problems for parallel analysis. With the help of MapReduce programming paradigm, which abstracts a parallel program that provides simplicity and flexibility to employ large-scale applications, MapReduce procedure of DIBID is developed by the following six steps (see Fig. 2).

1. Initial: In the beginning, the segmented original data in the independent Hadoop Distributed File System (HDFS) blocks are divided into training data and test data after data pre-processing. Then, the minority and majority classes in training data are separated.
2. Clustering: Clustering is applied independently to majority class and minority class for the detection of small disjuncts after determining the ideal cluster number.
3. Repartition: Each cluster is repartitioned in the most optimal number, which is roughly the same as the number of clusters.

TABLE I
DETAILS OF DATASETS USED IN EXPERIMENTS

Dataset	#Sample	Class (maj; min)	#Class (maj; min)	IR (maj/min)
Hepmass	10500000	(background; signal)	(5249876; 5250124)	0.9999
kdd_dos_normal	4856151	(dos; normal)	(3883370; 972781)	3.9920
kdd_dos_prb	3924472	(dos; prb)	(3883 370; 41102)	94.4812
kdd_dos_r2l	3884496	(dos; r2l)	(3883 370; 1126)	3 448.8188
kdd_dos_u2r	3883422	(dos; u2r)	(3883 370; 52)	74 680.1923
kdd_normal_prb	1013883	(normal; prb)	(972781; 41102)	23.6674
kdd_normal_r2l	973907	(normal; r2l)	(972781; 1126)	863.9262
kdd_normal_u2r	972883	(normal; u2r)	(972781; 52)	18 707.3269
skin	245057	(non-skin; skin)	(194198; 50859)	3.8183

4. Map: In clusters that are belong to the majority class, under-sampling is performed in a distributed manner in the direction of the <clusterId, Attributes> pairs, which is a <key, value> tuple. Similarly, over-sampling is performed in clusters of the minority class.
5. Reduce: After cluster-based resampling, the reduced majority class and the increased minority class are combined with the <classId, Attributes> pair.
6. Final: At this stage, the final data set is obtained to be used as a training set for classification. The size of this data varies according to the over-sampling and under-sampling rates.

IV. Experimental Study

In this section, first the details of data sets, methods, approaches and experimental environment are clarified. Then, the classification success of DIBID is evaluated. The evaluation of the results is carried out within two strategies: demonstrating the effect of DIBID on datasets with different IR values and comparing DIBID with other proposed models.

A. Experimental Framework

In order to analyse the performance of DIBID, experiments are run around three datasets: HEPMASS, KDD Cup 1999, and Skin Segmentation, which have different sample sizes and different imbalance ratios. Since relatively small data sets are used in studies performing imbalanced big data analysis, these datasets are chosen in order to make comparisons between models. As the KDD Cup 1999 consists of five classes as normal, dos, r2l, u2r, and prb, several binary combinations of these classes are created by filtration for binary-class classification. Table I summarises characteristics of selected datasets, where number of samples (#Sample), class labels (Class), number of instances (#Class), and Imbalance Ratio (IR) are presented.

To reduce variability, 5-fold cross-validation partitioning scheme is used and the average results are gathered to evaluate the model performance. 70 % of samples in datasets are selected as a training set and the remaining samples is considered as a test set. For clustering, resampling, and classification, training sets are scaled and analysis is performed on them. At the end of the analysis, classification results are from test sets.

The ideal *k* value for *k*-means clustering is usually a local minimum in the WSSSE (Within Set Sum of Squared Errors) graph. For this purpose, *k*-means clustering is run for each ***k*** value between 2 and 75 for each training set and the appropriate values are detected according to the point where there is an elbow in the WSSSE graph.

The distance between classes, the closeness of class elements to other class boundary, or the presence of within-class imbalance directly affect the ability of resampling techniques. In order to better observe the effects of techniques on different data distributions, the resampling is carried out following nine scenarios with different experimental rates:

1. Base: Not resampled;
2. UO: Getting 90 % of majority with RUS and increasing minority 100 % with ROS;
3. US: Getting 90 % of majority with RUS and increasing minority 100 % with SMOTE;
4. UO2: Getting 90 % of majority with RUS and increasing minority 200 % with ROS;
5. US2: Getting 90 % of majority with RUS and increasing minority 200 % with SMOTE;
6. U2O: Getting 80 % of majority with RUS and increasing minority 100 % with ROS;
7. U2S: Getting 80 % of majority with RUS and increasing minority 100 % with SMOTE;
8. U2O2: Getting 80 % of majority with RUS and increasing minority 200 % with ROS;
9. U2S2: Getting 80 % of majority with RUS and increasing minority 200 % with SMOTE.

Resampling processes on clusters are executed via repartitioning to parallelise distributed data processing by reducing the network traffic. To overcome this scalability problem, all the elements of the clusters are assembled together and one or more clusters, according to their sizes, are distributed in partitions to analyse simultaneously. Clusters whose number of elements are increased or decreased are collected according to minority or majority class labels and a new training set is created.

Because of RF’s outstanding performance, selecting the parameters at high values leads to high achievements. In order to determine that success is not due to the classification technique but resampling, the RF parameter specifications is

TABLE II
AUC RESULTS COMPARED TO OTHER IMBALANCED BIG DATA SOLUTIONS

Datasets	DIBID	[14]	Chi-FRBCS-BigData [22]	Chi-FRBCS-BigDataCS [22]	MEMMOT [9]	MMMmOT [9]	CMEOT [9]
kdd_dos_normal	**0.9999**	0.9998	0.9992	0.9993	*	*	*
kdd_dos_prb	**0.9999**	0.9994	0.8636	0.9557	*	*	*
kdd_dos_r2l	**0.9999**	0.9981	0.9913	0.9999	*	*	*
kdd_dos_u2r	0.8749	**0.9875**	0.8464	0.9366	*	*	*
kdd_normal_prb	**0.9983**	*	0.8932	0.9681	*	*	*
kdd_normal_r2l	**0.9690**	*	0.5050	0.9616	*	*	*
kdd_normal_u2r	**0.7999**	*	0.5000	0.5000	*	*	*
skin	**0.9961**	*	*	*	0.979	0.983	0.984

*the dataset was not used in related study

kept simple: the number of trees: 20 and maximum depth of each tree: 15.

All experiments are performed on GAZİ BIDISEC [35] cluster with 6 nodes connected with 4 × 10 Gb Ethernet. Each node is composed of 2 × 18-Core 2.3 GHz Intel E5-2699 microprocessors, 8 × 16 GB DDR4 Memory, and 12 × 8 TB SAS Disks. Apache Spark's MLlib is used for the classifications on the new training set.

B. Experimental Results

At the beginning of the development and analysis, minority and majority classes in the training sets are separated and clustered with their own sub-behaviours. According to the proposed MapReduce paradigm, after the minority class is increased and the majority class is reduced, a new training set is obtained. Finally, the new training set is classified with RF and the results are evaluated under two different strategies.

The first strategy is designed to present the success of DIBID model on the data sets with different IRs. For this purpose, RUS is applied to HEPMASS for simulating class imbalance problem by keeping the majority class constant and reducing the minority class [36]. Then, datasets with different IRs are created as: HEPMASS_1, HEPMASS_2, HEPMASS_3, and HEPMASS_4. Table III introduces some information on these new synthetic datasets such as the number of samples belonging to the majority (#maj) and minority (#min) classes, IRs, and AUC values as a result of base RF classification. At the next stage, DIBID is run with nine resampling scenarios and new training data sets are obtained. After applying RF to these data sets, the classification success rates are evaluated comparatively. Considering the rates of reaching the highest value from the base value, the increase in AUC approximately was 10 % (0.7469 to 0.8259) for HEPMASS_2, 24 % (0.5526 to 0.6882) for HEPMASS_3, and 18 % (0.5 to 0.5948) for HEPMASS_4, as seen in Fig. 3.

TABLE III
DETAILS OF SYNTHETIC HEPMASS DATASETS

Dataset	#maj	#min	IR(maj/min)	AUC
Hepmass_ 1	5249876	5249876	1	0.8645
Hepmass_ 2	5249876	524987	10	0.7439
Hepmass_ 3	5249876	52498	100	0.5526
Hepmass_ 4	5249876	5249	1000	0.500

The second strategy is designed to present the superiority of DIBID. For this purpose, the results of imbalanced big data solutions in the literature, which have approximately similar criteria with DIBID, are presented comparatively. The compared solutions are summarised as follows; [14] is a windowing approach for evolutionary under-sampling, Chi-FRBCS-BigData [22] is a fuzzy rule generation method, Chi-FRBCS-BigDataCS [22] is a cost-sensitive version of the previous method, MEMMOT [9] and MMMmOT [9] are enhanced SMOTE methodology, lastly CMEOT [9] is a cluster based over-sampling technique. In accordance with the results given in Table III, DIBID produced better AUC values in seven out of eight cases. This situation shows inevitable trade-off between the complexity of the analysis model and the difficulty in classifying the data.

V. DISCUSSION

That more amount of data is needed to create more comprehensive and robust models is not always true if the data set has irregularities. In real-world big data problems, data irregularities create more difficulties due to the large amount,

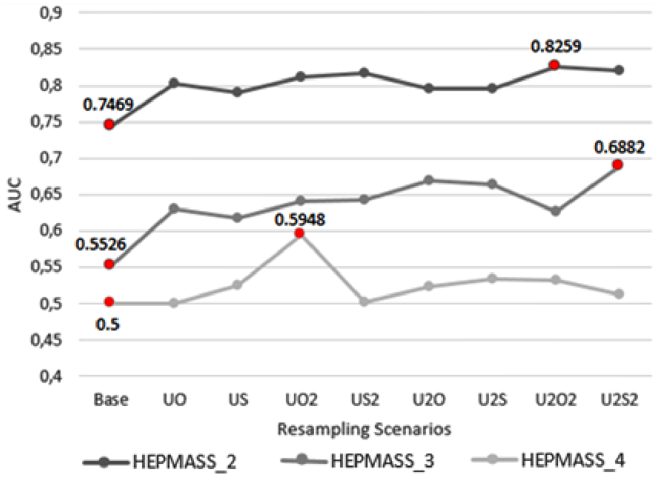

Fig. 3. AUC results of synthetic HEPMASS datasets on resampling scenarios.

high dimension and different sub-concepts. In addition, false negative can be costlier than false positive in real cases. While the DIBID experiments are performed to produce plausible solutions to the problem, various constraints and curative conditions are encountered. These issues can be summarised as follows:

- Data Types: High dimensional categorical data, which exists in many real-world problems, requires changes for distance function and representation of the centroids for neighbourhood finding and clustering. In order to realize the proof of concept, only numerical data is used in the present study;
- Cluster Number Selection: At this stage, *k* must be large enough to detect all small disjuncts. However, the lower WSSSEs, which occurs at big *k* values, may result in the possibility of almost each sample to be found in separate clusters. It is necessary to establish a good balance between these conditions. Another issue arises from the fact that repartition is performed according to the cluster number. Therefore, a cluster must be sized to fit at most one partition so that neighbourhoods do not remain in another partition;
- Partition Number Selection: Partitions provide parallelised distributed data processing with minimal network traffic. Accordingly, the workload in each partition should be approximately equal. For optimisation, small clusters may be placed in the same partition via repartitioning;
- IR Degree: In cases where the number of samples in classes is close to each other, resampling may cause the displacement of majority and minority classes. The proposed model is suitable for data sets with high IR value to avoid over-increment or over-reduction;
- Outlier Detection: Single-element clusters may occur even in an ideal clustering. This situation is considered as outlier and these clusters are filtered and not included in resampling within the study to avoid producing poor quality data;
- Memory Overhead: Due to the nature of most computations, analyses may be bottlenecked by CPU or memory. Although it is a challenging task to tune appropriate configuration in the large parameter space and the complex interactions among the parameters, it is mandatory to process large volume of data efficiently.

VI. Conclusion

In the present paper, new big data analytics based resampling model, named as DIBID, has been proposed for better classification results. MapReduce design of the model has been created and experiments performed within several scenarios. According to the results, the proposed DIBID outperformed other imbalanced big data solutions in the literature and increased AUC values between 10 % and 24 % through the case study.

Even if better and more robust results are achieved, a clearer and deeper conception is still needed for detecting class distribution impact on the learning process. Since big data problems have domain-specific nature, exploring idiographic solutions is very valuable. For these reasons, a better data understanding and more knowledge on the domain will be helpful in the analysis.

Acknowledgment

The authors wish to thank Gazi University Project Unit (GAZİ BAP) for its contribution to the Gazi University Big Data and Information Security Center (GAZİ BIDISEC) under grant 06/2015-04.

References

[1] M. K. Saggi and S. Jain, “A Survey Towards an Integration of Big Data Analytics to Big Insights for Value-Creation,” *Information Processing & Management*, vol. 54, no. 5, pp. 758–790, Sep. 2018. https://doi.org/10.1016/j.ipm.2018.01.010

[2] A. Oussous, F. Z. Benjelloun, A. A. Lahcen, and S. Belfkih, “Big Data Technologies: A survey,” *Journal of King Saud University – Computer and Information Sciences*, vol. 30, no. 4, pp. 431–448, Oct. 2018. https://doi.org/10.1016/j.jksuci.2017.06.001

[3] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, “Learning From Class-Imbalanced Data: Review of Methods and Applications,” *Expert Systems with Applications*, vol. 73, pp. 220–239, May 2017. https://doi.org/10.1016/j.eswa.2016.12.035

[4] H. He and E. A. Garcia, “Learning From Imbalanced Data,” *IEEE Transactions on Knowledge & Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009. https://doi.org/10.1109/TKDE.2008.239

[5] S. Das, S. Datta, and B. B. Chaudhuri, “Handling Data Irregularities in Classification: Foundations, Trends, and Future Challenges,” *Pattern Recognition*, vol. 81, pp. 674–693, Sep. 2018. https://doi.org/10.1016/j.patcog.2018.03.008

[6] J. Stefanowski, “Dealing With Data Difficulty Factors While Learning From Imbalanced Data,” in *Challenges in Computational Statistics and Data Mining*, pp. 333–363, 2016. https://doi.org/10.1007/978-3-319-18781-5_17

[7] A. Fernández, S. del Río, N. V. Chawla, and F. Herrera, “An Insight Into Imbalanced Big Data Classification: Outcomes and Challenges,” *Complex & Intelligent Systems*, vol. 3, no. 2, pp. 105–120, Jun. 2017. https://doi.org/10.1007/s40747-017-0037-9

[8] S. del Río, V. López, J. M. Benítez, and F. Herrera, “On the Use of MapReduce for Imbalanced Big Data Using Random Forest,” *Information Sciences*, vol. 285, pp. 112–137, 2014. https://doi.org/10.1016/j.ins.2014.03.043

[9] S. S. Patil and S. P. Sonavane, “Enriched Over_Sampling Techniques for Improving Classification of Imbalanced Big Data,” in *2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService)*, USA, 2017, pp. 1–10. https://doi.org/10.1109/BigDataService.2017.19

[10] M. Ghanavati, R. K. Wong, F. Chen, Y. Wang, and C. S. Perng, “An Effective Integrated Method for Learning Big Imbalanced Data,” in *2014 IEEE International Congress on Big Data*, USA, 2014, pp. 691–698. https://doi.org/10.1109/BigData.Congress.2014.102

[11] D. Galpert, S. del Río, F. Herrera, E. Ancede-Gallardo, A. Antunes, and G. Agüero-Chapin, “An Effective Big Data Supervised Imbalanced Classification Approach for Ortholog Detection in Related Yeast Species,” *BioMed Research International*, vol. 2015, Article ID 748681, 2015. https://doi.org/10.1155/2015/748681

[12] S. del Río, J. M. Benítez, and F. Herrera, “Analysis of Data Preprocessing Increasing the Oversampling Ratio for Extremely Imbalanced Big Data Classification,” in *2015 IEEE Trustcom/BigDataSE/ISPA*, pp. 180–185, Finland, 2015. https://doi.org/10.1109/Trustcom.2015.579

[13] I. Triguero, S. del Río, V. López, J. Bacardit, J. M. Benítez, and F. Herrera, “ROSEFW-RF: The Winner Algorithm for the ECBDL’14 Big Data Competition: An Extremely Imbalanced Big Data Bioinformatics Problem,” *Knowledge-Based Systems*, vol. 87, pp. 69–79, Oct. 2015. https://doi.org/10.1016/j.knosys.2015.05.027

[14] I. Triguero, M. Galar, S. Vluymans, C. Cornelis, H. Bustince, F. Herrera, and Y. Saeys, “Evolutionary Undersampling for Imbalanced Big Data Classification,” in *2015 IEEE Congress on Evolutionary Computation (CEC)*, Japan, 2015, pp. 715–722. https://doi.org/10.1109/CEC.2015.7256961

[15] I. Triguero, M. Galar, D. Merino, J. Maillo, H. Bustince, and F. Herrera, “Evolutionary Undersampling for Extremely Imbalanced Big Data Classification Under Apache Spark,” in *2016 IEEE Congress on Evolutionary Computation (CEC)*, Canada, 2016, pp. 640–647. https://doi.org/10.1109/CEC.2016.7743853

[16] S. Kamal, S.H. Ripon, N. Dey, A.S. Ashour, and V. Santhi, “A MapReduce approach to diminish imbalance parameters for big deoxyribonucleic acid dataset,” *Computer methods and programs in biomedicine*, vol. 131, pp. 191–206, Jul. 2016. https://doi.org/10.1016/j.cmpb.2016.04.005

[17] F. Hu, H. Li, H. Lou, and J. Dai, “A parallel oversampling algorithm based on NRSBoundary-SMOTE,” *Journal of Information & Computational Science*, vol. 11, no. 13, pp. 4655–4665, Sep. 2014. https://doi.org/10.12733/jics20104484

[18] R. C. Bhagat and S. S. Patil, “Enhanced SMOTE Algorithm for Classification of Imbalanced Big-Data Using Random Forest,” in *2015 IEEE International Advance Computing Conference (IACC)*, India, 2015, pp. 403–408. https://doi.org/10.1109/IADCC.2015.7154739

[19] C. K. Maurya, D. Toshniwal, and G. V. Venkoparao, “Online Sparse Class Imbalance Learning on Big Data,” *Neurocomputing*, vol. 216, pp. 250–260, Dec. 2016. https://doi.org/10.1016/j.neucom.2016.07.040

[20] M. Tang, C. Yang, K. Zhang, Q. Xie, “Cost-Sensitive Support Vector Machine Using Randomized Dual Coordinate Descent Method for Big Class-Imbalanced Data Classification,” *Abstract and Applied Analysis*, vol. 2014, Article ID 416591, Jul. 2014. https://doi.org/10.1155/2014/416591

[21] X. Wang, X., Liu, and S. Matwin, “A distributed instance-weighted SVM algorithm on large-scale imbalanced datasets”. in *2014 IEEE International Conference on Big Data*, USA, 2014, pp. 45–51. https://doi.org/10.1109/BigData.2014.7004467

[22] V. López, S. del Río, J. M. Benítez, and F. Herrera, “Cost-Sensitive Linguistic Fuzzy Rule Based Classification Systems Under the MapReduce Framework for Imbalanced Big Data,” *Fuzzy Sets and Systems*, vol. 258, pp. 5–38, Jan. 2015. https://doi.org/10.1016/j.fss.2014.01.015

[23] S. del Rio, V. Lopez, J. M. Benítez, and F. Herrera, “A MapReduce Approach to Address Big Data Classification Problems Based on the Fusion of Linguistic Fuzzy Rules,” *International Journal of Computational Intelligence Systems*, vol. 8, no. 3, pp. 422–437, May 2015. https://doi.org/10.1080/18756891.2015.1017377

[24] J. Zhai, S. Zhang, M. Zhang, and X. Liu, “Fuzzy Integral-Based ELM Ensemble for Imbalanced Big Data Classification,” *Soft Computing*, vol. 22, no. 11, pp. 3519–3531, Jun. 2018. https://doi.org/10.1007/s00500-018-3085-1

[25] Z. Wang, J. Xin, H. Yang, S. Tian, G. Yu, C. Xu, and Y. Yao, “Distributed and Weighted Extreme Learning Machine for Imbalanced Big Data Learning,” *Tsinghua Science and Technology*, vol. 22, no. 2, pp. 160–173, Apr. 2017. https://doi.org/10.23919/TST.2017.7889638

[26] N. B. Abdel-Hamid, S. ElGhamrawy, A. El Desouky, and H. Arafat, “A Dynamic Spark-Based Classification Framework for Imbalanced Big Data,” *Journal of Grid Computing*, vol. 16, no. 4, pp. 607–626, Dec. 2018. https://doi.org/10.1007/s10723-018-9465-z

[27] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, “A Survey on Addressing High-Class Imbalance in Big Data,” *Journal of Big Data*, vol. 5, no. 42, Dec. 2018. https://doi.org/10.1186/s40537-018-0151-6

[28] J. W. Huang, C. W. Chiang, and J. W. Chang, “Email Security Level Classification of Imbalanced Data Using Artificial Neural Network: The Real Case in a World-Leading Enterprise,” *Engineering Applications of Artificial Intelligence*, vol. 75, pp. 11–21, Oct. 2018. https://doi.org/10.1016/j.engappai.2018.07.010

[29] T. Jo, and N. Japkowicz, “Class Imbalances Versus Small Disjuncts,” *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 40–49, Jun. 2004. https://doi.org/10.1145/1007730.1007737

[30] A. Agrawal, H. L. Viktor, E. Paquet, “SCUT: Multi-Class Imbalanced Data Classification Using SMOTE and Cluster-Based Undersampling,” in *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, 2015, vol. 1, pp. 226–234. https://doi.org/10.5220/0005595502260234

[31] W. C. Lin, C. F. Tsai, Y. H. Hu, and J. S. Jhang, “Clustering-Based Undersampling in Class-Imbalanced Data,” *Information Sciences*, vol. 409, pp. 17–26, Oct. 2017. https://doi.org/10.1016/j.ins.2017.05.008

[32] I. Nekooeimehr and S. K. Lai-Yuen, “Adaptive Semi-Unsupervised Weighted Oversampling (A-SUWO) for Imbalanced Datasets,” *Expert Systems with Applications*, vol. 46, pp. 405–416, Mar. 2016. https://doi.org/10.1016/j.eswa.2015.10.031

[33] A. Estabrooks, T. Jo, and N. Japkowicz, “A Multiple Resampling Method for Learning from Imbalanced Data Sets,” *Computational Intelligence*, vol. 20, no. 1, pp. 18–36, Feb. 2004. https://doi.org/10.1111/j.0824-7935.2004.t01-1-00228.x

[34] H. Guo, J. Zhou, and C. A. Wu, “Imbalanced Learning Based on Data-Partition and SMOTE,” *Information*, vol. 9, no. 238, Sep. 2018. https://doi.org/10.3390/info9090238

[35] GAZİ-BIDISEC. Gazi University Big Data and Information Security Center. [Online]. Available: http://bigdatacenter.gazi.edu.tr/ [Accessed: Sep. 2019].

[36] T. Hasanin and T. Khoshgoftaar, “The Effects of Random Undersampling with Simulated Class Imbalance for Big Data,” in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, USA, 2018, pp. 70–79. https://doi.org/10.1109/IRI.2018.00018

Duygu Sinanc Terzi is a Research Assistant and a Doctoral student at the Department of Computer Engineering of Gazi University, Ankara, Turkey. Her research interests include big data analytics, machine learning, information security, and anomaly detection.
E-mail: duygusinanc@gazi.edu.tr
ORCID ID: https://orcid.org/0000-0002-3332-9414

Seref Sagiroglu is a Professor at the Department of Computer Engineering of Gazi University, Ankara, Turkey. His research interests include intelligent system identification, recognition, modelling, and control; artificial intelligence; heuristic algorithms; industrial robots; information systems and applications; software engineering; information and computer security; biometry; big data analytics; malware and spyware software.
E-mail: ss@gazi.edu.tr
ORCID ID: https://orcid.org/0000-0003-0805-5818