ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 8, August 2018

Parallel Heterogeneous Voting Ensemble for Effective Classification of Imbalanced Data

K.Madasamy[1], Dr.M.Ramaswami[2]

Research Scholar (Part-Time), Department of Computer Applications, Madurai Kamaraj University, Madurai, India[1]

Professor, Department of Computer Applications, Madurai Kamaraj University, Madurai, India[2]

ABSTRACT: In recent times, data deluge is an inevitable phenomenon that emerges from modern electronic gadgets that we use in our day-to-day life. Knowledge extraction from these voluminous data is a complex task. In particular, handling massive imbalanced data is one of the major bottlenecks in the classification process. At present, classification of imbalanced data has gained much attention in all domains of research. This paper proposes a heterogeneous ensemble model for effective classification of imbalanced data. This model uses multiple tree based base-learners for predictions. Each of the base-learners has its own pros and cons. The base-learners are chosen in such a way that each one complement the other one and hence providing enhanced predictions. Voting based combiner is used to aggregate results. The aforesaid model performs both binary and multi-class classifications. Apache Spark based implementation enables the model to handle big data effectively. Experiments and comparisons on existing state-of-the-art models reveal effective performances and exhibiting the increased improvements in terms of AUC levels.

KEYWORDS: Data Imbalance, Bagging, Heterogeneous Models, Decision Tree, Random Forest, Gradient Boosted Trees, RHSBoost.

I. INTRODUCTION

Data Imbalance is a classical problem in real-time based applications that involve domains concerning rare occurrences. Examples include fraud detection in credit cards [29], anomaly detection in network traffic data [33], medical diagnostics in detecting cancer cells and distinguishing them from normal cells [25], etc. The common occurrence is that normal data is high in density, while the data of interest is low in density, sometimes very low, creating imbalance. The imbalance nature of data disrupts the prediction process. This necessitates specific learning techniques called imbalance learning.Machine learning models are usually used for classification. However, machine learning algorithms usually consider their input data to be balanced. Hence data with imbalance requires specialized models that concentrate on the minority classes (classes with low density) [2]. Although several applications in the imbalanced domain require binary classifiers, some special cases like document classification require multi-class classification models [21]. There exist two major ways to create models specialized in imbalanced learning. The first is to modify the objective function of the model to concentrate more on the minority classes, the second is to resample the data to reduce imbalance. Resampling is performed by over-sampling the minority classes, under-sampling the majority classes or by hybrid sampling [8]. This work proposes a heterogeneous ensemble with voting as the combiner element to provide an effective model for classifying imbalanced data.

The remainder of this paper is structured as follows; section II presents the insight on review of literature, section III presents a detailed discussion of the proposed heterogeneous ensemble model, section IV presents the results and section V concludes the work.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 8, August 2018

II. RELATED WORK

Increased automation has led to high data generation and requirements to capture and process this data. All these real-time data are usually laden with imbalances. Hence recent years have seen a huge increase in the requirements for handling imbalanced data. This section discusses some of the recent contributions in this domain.

A boosting based model was proposed by Gong et al. [14]. This model is a sampling based method that uses multiple sampling models to achieve the desired balance. This model also proposes a boosting technique for effective prediction of classes. A comparison of multiple ensemble models for imbalance data prediction was proposed by Galar et al. [13]. An imbalanced data classifier for binary classification was proposed by Du et al. [10]. This model is based for Neural Networks and claims to significantly improve the performance of neural networks. This model initially creates classifier rules using neural networks and refines them using geometric mean.

A bagged ensemble specifically designed for credit card fraud detection was proposed by Akila et al. [1]. This model proposes a bagging methodology for effective detection of fraudulent cases in credit card transactions. A cost sensitive model to handle imbalance was proposed by Liu et al. [23]. This is a probability estimation based classifier model, aimed to effectively handle data imbalance. A credit classification method to handle imbalanced data was proposed by Yu et al. [35]. This is a rebalancing mechanism that utilizes bagging and resampling models to perform predictions. A study on imbalanced data and metrics to measure their performances was proposed by Borsos et al. [5]. A dissimilarity based imbalance handling model was proposed by Zhang et al. [36]. This work concentrates on feature selection and constructs a dissimilarity space to provide effective predictions.A comparative study on performance of different classifiers for multiclass problem was carried in [31]. In this study, the use of ROC curves is suggested as the best tool for visualizing and analysing the behaviour of various classifiers.

An under sampling model to handle imbalance was proposed by Triguero et al. in [32]. This model proposes a classification algorithm and claims to effectively operate on data with high imbalance levels. A spark based Random Forest model was proposed by Chen et al. in [9]. This model creates a parallelized Random Forest algorithm for effective classification in Spark environment. Other spark based models include SCARFF [7] and an ensemble model [34].

Alberto Fernandez etal. [3] has conducted a study on imbalanced big data classification and its varied challenges; they proposed the behaviour of some of the standard existing pre-processing techniques using Spark library to adopt with big data domain. Further they analysed the performance achieved by combining the different classification algorithms with varied pre-processing techniques in the context of imbalanced big data problems. Mehdi Assefi etal. [26] have proposed comparative study of various classifier models using SVM, Random Forest, Decision Tree and Nave Bayes in terms of Area Under ROC and execution time with Apache Spark MLib and Weka library. The comparative study reveals that the Apache Spark MLib is performing faster than Weka. NeelamNaik and SeemaPurohit [28] have conducted a comparative study on binary classification method using Decision Tree, Random Forest Tree and Gradient Boosted Tree classifiers. The experiments were conducted using Cloud based Apache Spark environment; the authors demonstrates that random forest is the best among all three trees and it takes minimum time for building the classification tree. JalaSri etal. [19], have proposed an efficient DS-RNGE nearest neighbour classification solution to classify high speed, large-scale data streams using Apache Spark. Jorge Gonzalez-Lopez etal. [20], have conducted an experimental study on large-scale multi-label ensemble learning under Apache Spark environment. The authors have used five combinations of the parallel and distributed approaches like Mulan, Mulan Threading, Mulan distributed, Mulan distributed threading and Spark, over 29 multi-label datasets collected from MULAN, MEKA and LABIC repositories websites. The authors concluded that native Spark implementation outperforms the other remaining approaches and it is the solid framework capable of distributing the workload of large processing tasks like multi-label ensembles.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 8, August 2018

III.HETEROGENEOUS VOTING ENSEMBLE

Data imbalance is identified to be a common occurrence in data generated in real-time. Although the imbalance levels vary between domains, their presence cannot be overlooked, due to the major performance issues they impose upon classifier models. Certain levels of imbalance are automatically handled by classifiers, however, the actual levels of imbalance automatically handled by the classifiers are dependent on the data distribution and the classifier's ability, and hence it varies with data and models. Imbalance in data still poses a huge threat by taxing on the classifier's performance if left unchecked, hence handling imbalance in data still remains a huge challenge.This work proposes an ensemble based classifier that incorporates heterogeneous ensembles to perform classification and a voting combiner to combine results (Figure 1). The proposed architecture is composed of three major sections; the data preparation phase, classification process and the voting combiner.

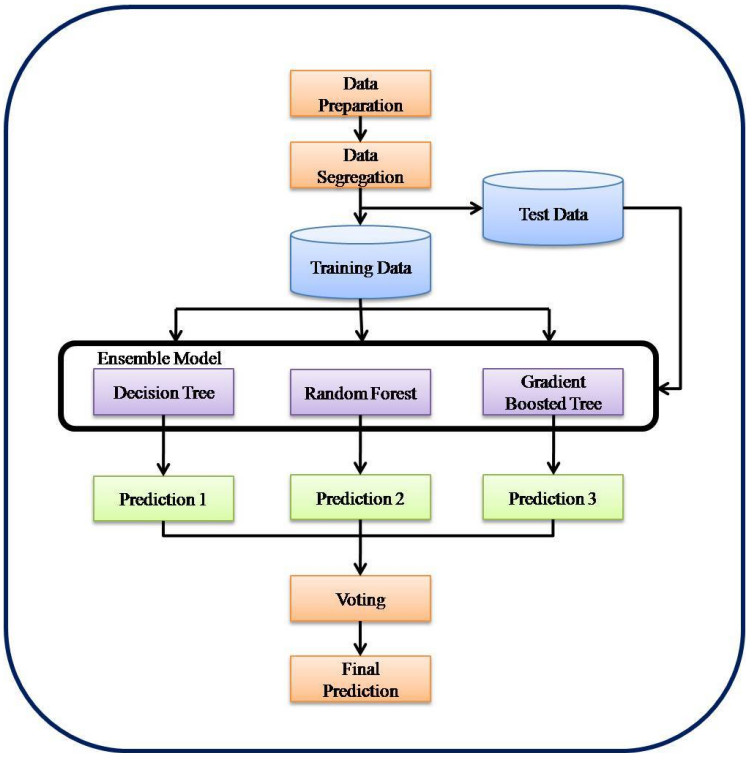

Fig 1.The proposed Heterogeneous Voting Ensemble Framework

Data Preparation

Data preparation is one of the major components of any rule building model. This is due to the fact that qualities of the rules have a major dependency on the quality of the data being used for the rule building process. Further, most of the existing classifier models operate only on clean and consistent data. This mandates the need for a data preparation phase. Data preparation aims to eliminate the inconsistencies contained in the data. However, eliminating the inconsistencies require identification of such instances. Hence data analysis forms a major phase in this process. Data analysis aims to identify missing data and erroneous data. The erroneous data is corrected and missing values are filled with appropriate data to obtain the clean data. The cleaned data is then segregated to obtain the training data and the

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 8, August 2018

test data in the ratio 7:3, i.e. 70% of the available data is used for training and 30% of the available data is used for testing.

Heterogeneous Multi-Model Based Classification

We useheterogeneous based multi-modelclassification phase in the proposed work. Training data obtained from the previous phase is used as the input to this phase. This work utilizes multiple classifiers for prediction, rather than multiple instances of a single classifier. This introduces heterogeneity into the prediction process, hence reducing the bias and variance in the prediction process. The proposed work uses tree based models for creating the ensemble. The classifier models used for the process are, Decision Tree, Random Forest and Gradient Boosted Trees. Although all the three models are tree based, only Decision Trees are generic and the other two models are formed by Bagging and Boosting Decision Trees, hence their performances were observed to considerably vary from Decision Trees. The nature of this modelling was introduced, such that the downside of one model is complemented by another model, thereby improving the overall prediction efficiency.

Decision Tree [24,30] is a graph based model, built using graph like structures with nodes and leaves. Decision Trees represent decisions and their corresponding consequences. A decision tree is composed of a root, multiple nodes and leaves. The root node is usually the beginning point. All nodes represent conditions and all leaves represent classes or final decisions. Decision rules are obtained by following the constructed tree from the root to the leaf according to the conditions. The major advantage of this model is that it is a simple and white-box model and can be easy to walk-through. Further, it is dynamic and hence conditions can be incorporated into the model during run-time. The major disadvantage of this model is that it is a weak classifier, hence, as the underlying data changes; a huge change in the structure is required. Further, categorical variables can have a huge negative impact on the model, such that categorical variables with low occurrence frequencies tend to go under-represented. They also tend to over-fit to the data instances.

Random Forest [15,16] is a homogeneous ensemble prediction model, created by incorporating multiple Decision Trees as base learners. Random Forest is a bagging based ensemble, created using multiple decision trees and passing a subset of data to each of these base learners for training. The combiner provides the final results. The major advantage of Random Forest is that it provides consistency to the model, and hence provides a strong classifier. It tends to solve the over-fitting issue contained in the Decision Trees.

Gradient Boosted Trees [6,11,12] are a class of ensemble models, created by applying the Boosting technique on a base learner. Boosting operates by iteratively solving the problem by incorporating the error component into the training process, thereby reducing the error in the prediction process. The major advantage of this model is that it helps in generalization and reduces variance in the training process, however, it tends to increase bias in the model.

The entire training data obtained from the previous phase is passed to the constructed ensemble model. Each base learner builds classification rules based on the training data. The data to be predicted is passed to the constructed ensemble model, and each model provides its own prediction for the given data. A single instance in the data now contains three predictions, a prediction from each base learner.

Voting

Multiple predictions obtained from the previous phase are passed as input to this phase. Although predictions are performed on the same instances, a prediction result from each model varies considerably. This variance observed in the prediction process is due to the short-comings observed in each model. Hence in-order to obtain a single stable result, this work uses a voting combiner. Voting is the process of combining results such that majority predictions are considered as the final predictions. Voting is the usual grouping mechanism used in bagged ensembles. The result obtained from the voting combiner is considered as the final predictions.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 8, August 2018

IV. RESULTS AND DISCUSSION

The proposed heterogeneous ensemble model has been coded in PySpark to enable the model to be operated upon big data. PySpark is the Python API for working with Apache Spark [18] environment. The proposed model is being implemented in Databricks [17] community cloud platform. This community edition provides single cluster with up to 6 GB free storage. Databricks is an integrated and sophisticated data analytics platform created by Apache Spark which aids to deploy machine learning models over the Amazon cloud in a seamless way.

Five data sets with varied imbalance levels ranging from 13 to 115 were chosen from UCI [22, 27] and KEEL repository [4] for analysis. Both binary and multi class datasets are chosen to analyse the efficiency of the model on datasets of varied dimensions and class labels.Description for all the datasets are shown in Table1.

Table1: Dataset Description

Name of the Dataset	No.of Instances	No.of Attributes	Imbalance Ratio	No. of Classes	Source
Cover Type	38501	10	13.02	Multi (7)	UCI
Glass5	214	9	22.81	Binary (2)	KEEL
Wine	4898	11	25.77	Multi (3)	UCI
Yeast6	1484	8	39.15	Binary (2)	KEEL
Abalone	4177	8	115.03	Binary (2)	UCI

A table depicting the performance levels of the proposed model on all the datasets is shown in Table 2. It could be observed that the proposed model yields increased performance levels in terms of the metrics.

Table 2: Performance Metrics

Data	AUC	Accuracy	F1-Score	Recall	Precision
CoverType	0.9713	0.9292	0.9139	0.9202	0.9077
Glass5	1	1	1	1	1
Wine	0.9958	0.9693	0.9608	0.9636	0.9580
Yeast6	0.9690	0.9977	0.9984	0.9980	0.9986
Abalone	1	1	1	1	1

AUC of the proposed classifier, representing the area under the curve is shown in figure 2. It could be observed that the proposed model shows an AUC level greater than 0.9 for all the models. AUC is considered to be a balanced measure, providing the aggregated results depicting the true prediction levels and false positive levels of a classifier model. High true prediction levels and low false prediction levels reveals the characteristics of a good classifier.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 8, August 2018

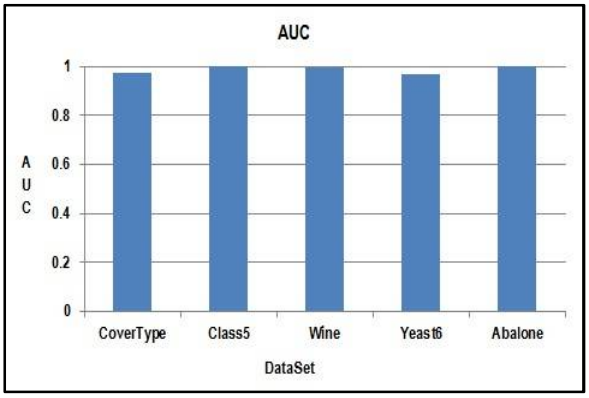

Fig.2 AUC

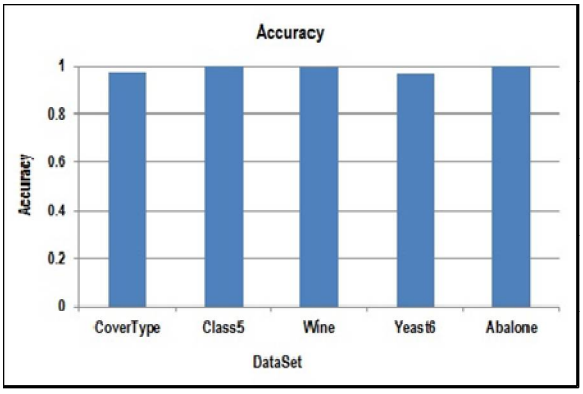

Fig.3 Accuracy

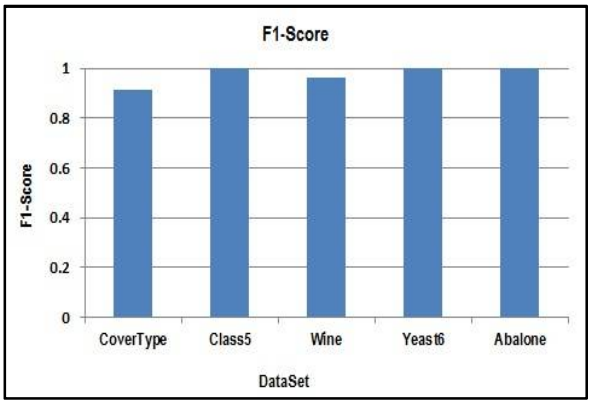

Fig. 4 F1-score

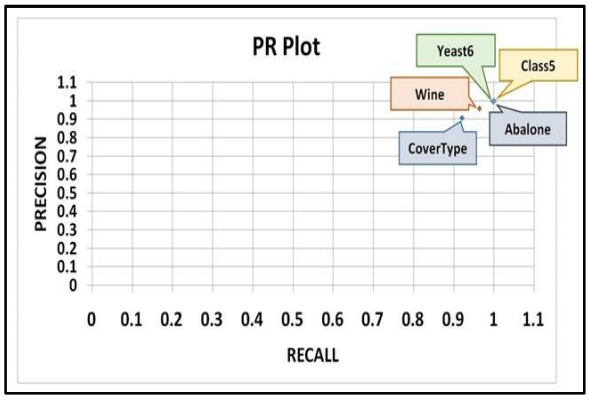

Fig.5 PR Plot

Accuracy levels obtained for the proposed model on the selected datasets is shown in figure 3. It is clear that the accuracy levels obtained for all the datasets are greater than 0.9 depicting the high performing nature of the proposed model.

F1-Score is another aggregate measure showing the overall performance of a classifier in terms of its precision and recall levels. It becomes visible from figure 4 that the proposed model produces F1-Scores greater than 0.9 which shows the high performing nature of the heterogeneous ensemble model.

The PR plot of the proposed model is shown infigure 5. It could be observed that the proposed model yields high precision and recall levels irrespective of the imbalance level contained in the data. Recall refers to the ability of the model to retrieve relevant results, while precision refers to level of relevancy obtained in the results. Both the measures are expected to be high in a classifier, as both the metrics are equally important in measuring the performance of a classifier.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 8, August 2018

A comparison of the proposed model with RHSBoost [14], a boosting based ensemble, and two phase stacking model [24] is shown in Figure 6. The models are compared in terms of AUC levels. It is to be confirmed that, the proposed model indicatesincreased performance in all the datasets when compared with both the existing models. Although Glass5 data is an exception to this, the variance is very low (~0.01), hence is considered trivial.

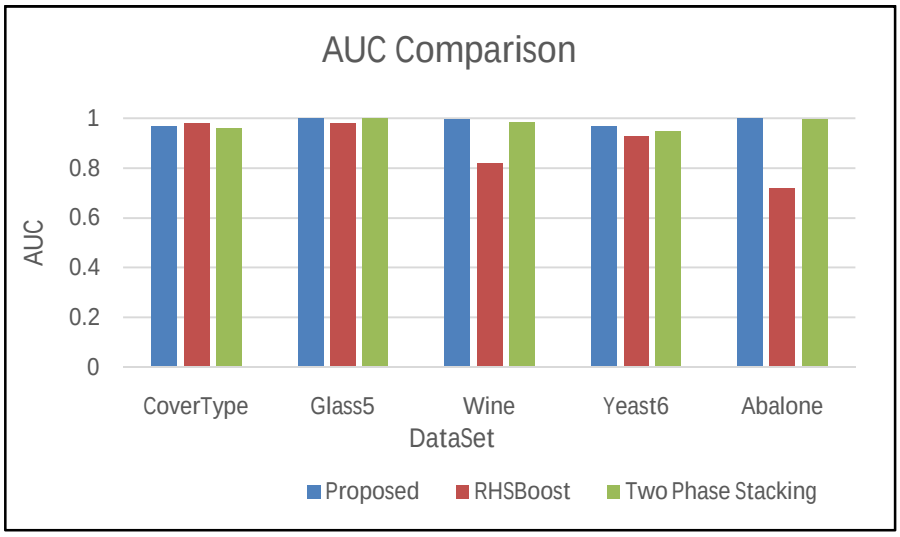

Fig.6 AUC Comparison with RHSBoost and Two Phase Stacking methods

V. CONCLUSION

Imbalanced data is one of the major issues in practice that complicates the classification process. Although multiple models exist for handling imbalanced data, their usage domains are limited. Most of the models are used for specific purposes and they cannot be generalized. This is the reason why there are lacks of models for general data analysis when imbalanced data is concerned. The model specified in this research proposes an effective heterogeneous ensemble for classification of imbalanced data. The proposed model has been implemented in Apache Spark to handle huge data. And this model is generic, scalable and can be used for both binary and multi-class classification. The base learners are chosen to enable effective predictions in such a way that the negative of one base learner is adjusted by the other base learners. Experiments were performed with standard benchmark datasets and comparisons with the existing models indicate increased performance levels in terms of Precision, Recall, F-Measure and AUC. However, the proposed model has its own limitation of using it only with low level to moderate level of imbalanced datasets. It is suggested that either extension of this model or a new model may be developed for handling data with high level of imbalance.

REFERENCES

[1] Akila.S, andSrinivasuluReddy.U. "Risk based bagged ensemble (RBE) for credit card fraud detection." Inventive Computing and Informatics (ICICI), International Conference on. IEEE, 2017.
[2] Akila.S, and SrinivasuluReddy.U., "Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data." Proceedings of ICRECT volume .16,pp. 28-34, 2016.
[3] Alberto Fernandez, Sara del Rio, NiteshV.Chawla, Francisco Herrera,(2017) “An insight into imbalanced Big Data classification: outcomes and challenges”, complex Intell. Syst. DOI 10.1007/s40747-017-0037-9.
[4] Alcalá-Fdez, Jesús, et al. "Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework." Journal of Multiple-Valued Logic & Soft Computing, Vol 17, 2011.
[5] Borsos, Zalán, CameliaLemnaru, and RodicaPotolea. "Dealing with overlap and imbalance: a new metric and approach." Pattern Analysis and Applications, pp. 1-15.2016.
[6] Breiman, L. "Arcing The Edge" June 1997.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 8, August 2018

[7] Carcillo, F., Dal Pozzolo, A., Le Borgne, Y.A., Caelen, O., Mazzer, Y. and Bontempi, G., "SCARFF: A scalable framework for streaming credit card fraud detection with spark". Information fusion, 41, pp.182-194, 2018.

[8] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research volume. 16: pp 321-357 2002.

[9] Chen, J., Li, K., Tang, Z., Bilal, K., Yu, S., Weng, C. and Li, K., "A parallel random forest algorithm for big data in a spark cloud computing environment". IEEE Transactions on Parallel & Distributed Systems, (1), pp.1-1, 2017.

[10] Du, Jie, et al. "Post-boosting of classification boundary for imbalanced data using geometric mean." Neural Networks, volume. 96 pp. 101-114. 2017.

[11] Friedman, J. "Greedy Function Approximation: A Gradient Boosting Machine http://www. salford-systems. com/doc.", GreedyFuncApproxSS. Pdf, 1999.

[12] Friedman, J. H. "Stochastic Gradient Boosting." March 1999.

[13] Galar.M, Fernandez.A, Barrenechea.E, Bustince.H, &Herrera.F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), volume. 42(4), pp. 463-484. 2012.

[14] Gong Joonho, and Hyunjoong Kim. "RHSBoost: Improving classification performance in imbalance data." Computational Statistics & Data Analysis, volume. 111 pp. 1-13.2017.

[15] Ho, Tin Kam. "Random decision forests." Document analysis and recognition, 1995., proceedings of the third international conference on. volume. 1. IEEE, 1995.

[16] Ho, Tin Kam. "The random subspace method for constructing decision forests." IEEE transactions on pattern analysis and machine intelligence volume. 20 no 8 pp. 832-844. 1998.

[17] https://databricks.com/

[18] https://spark.apache.org/

[19] Jalasri.M, Aarthika.K, Vishnu Priya.A, "High Speed Classification of Massive Data Streaming Using Spark", International Journal of Advance Engineering and Research Development, Vol-5, Issue 02, Feb 2018.

[20] Jorge Gonzalez-Lopez, Alberto Cano, Sebastian Ventura, "Large-scale multi-label ensemble learning on Spark", DOI 10.1109/Trustcom/BigDataSE/ICESS.2017.328.

[21] Kumar, M. Arun, and MadanGopal. "Reduced one-against-all method for multiclass SVM classification." Expert Systems with Applications volume.38, no.11 pp.14238-14248. 2011.

[22] Lichman, Moshe. "UCI machine learning repository." 2013.

[23] Liu, Zhenbing, et al. "Cost-Sensitive Collaborative Representation Based Classification via Probability Estimation Addressing the Class Imbalance Problem." Artificial Intelligence and Robotics. Springer, Cham, pp. 287-294. 2018.

[24] Madasamy.K, and Ramaswami.M, "Two-Phase Stacking Ensemble To Effectively Handle Data Imbalances In Classification Problems." International Journal of Advanced Research in Computer Science volume 9, no. 1: pp. 908-913. 2018.

[25] Mazurowski.M.A. ,Habas.P.A. ,Zurada.J.M.,, Lo., J.Y., Baker. J.A, and Tourassi.G.D, . "Training neural network classifiers for medical decision making". The effects of imbalanced datasets on classification performance. Neural networks, volume 21(2-3), pp.427- 436. 2008.

[26] Mehdi Assefi, EhsunBehravesh, Guangchi Liu and Ahmad P.Tafti, "Big Data Machine Learning using Apache Spark MLlib, IEEE Big Data 2017.

[27] Menardi.G, &Torelli.N, Training and assessing classification rules with imbalanced data. DataMining and Knowledge Discovery,volume 28(1), pp. 92-122.2014.

[28] NeelamNaik, SeemaPurohit, "Comparative Study of Binary Classification Methods to Analyze a Massive Dataset on Virtual Machine ", International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France.

[29] Phua, Clifton, DammindaAlahakoon, and Vincent Lee. "Minority report in fraud detection: classification of skewed data." Acmsigkdd explorations newsletter, volume. 6, no.1: pp50-59, 2004.

[30] Quinlan, J. Ross. "Simplifying decision trees." International journal of man-machine studies, volume 27 no.3,pp. 221-234. 1987.

[31] Ramaswami.M, "Validating Predictive Performance of Classifier Models for Multiclass Problem in Educational Data Mining", International Journal of Computer Science Issues, Vol.11, Issue 5, No.2, September 2014.

[32] Triguero, I., Galar, M., Merino, D., Maillo, J., Bustince, H. and Herrera, F., "Evolutionary undersampling for extremely imbalanced big data classification under apache spark". In Evolutionary Computation (CEC), 2016 IEEE Congress on (pp. 640-647). IEEE, 2016.

[33] Umamaheswari. K, and Janakiraman.S "Intrusion detection on large imbalanced data using hybridized firefly algorithm (HFA) ", pp. 1599-1612, 2017.

[34] Wu, Z., Lin, W., Zhang, Z., Wen, A. and Lin, L.,"An ensemble random forest algorithm for insurance big data analysis". In Computational Science and Engineering (CSE) and Embedded and Ubiquitous Computing (EUC), 2017 IEEE International Conference on (Vol. 1, pp. 531-536). IEEE,2017.

[35] Yu, Lean, et al. "A DBN-based resampling SVM ensemble learning paradigm for credit classification with imbalanced data." Applied Soft Computing, volume. 69, pp. 192-202.2018.

[36] Zhang, Xueying, et al. "A dissimilarity-based imbalance data classification algorithm." Applied Intelligence volume 42, no. 3, pp. 544-565, 2015.