E-ISSN 2581 – 7957
P-ISSN 2277 – 3916

CVR Journal of Science and Technology, Volume 17, December 2019
DOI: 10.32377/cvrjst1714

# Imbalanced Big Data Classification using Feature Selection Under-Sampling

Ch. Sarada[1] and M. Sathya Devi[2]
[1] Asst. Professor, CVR College of Engineering/CSE Department, Hyderabad, India.
Email: sharada.ch@gmail.com
[2] Asst. Professor, CVR College of Engineering/CSE Department, Hyderabad, India.
Email: satyamaranganti@gmail.com

***Abstract:*** **Imbalanced learning is the classification problem where the number of observations of one class, far surpasses the number of observations of another class. Different sampling approaches are proposed for paired and Multi-Class imbalanced classification. Paired Imbalanced classification encompasses two classes: one of them is majority, while the other one is a minority class. Multi-Class imbalanced classification contains more than two classes for classification. Under-sampling technique is the better sampling technique among conventional approaches. However, existing approaches may not work in the Big Data environment, as considering all the features might compromise the performance of the system. In this work, a novel method is presented which takes into account only the essential features, as well as, deals with massive data as in Big Data environment. In the proposed system, Feature Selection Under-Sampling technique is used for resampling the data. Feature selection is the vital step because it not only decreases the dimensionality of data but also helps classifier to run faster, and accuracy can also be improved. Over that, SVM learning classifier is adopted to construct the model and test the data. The proposed system is implemented using MapReduce framework by integrating statistical analytical tool R.**

***Index Terms:*** **Big Data, Imbalanced learning, Sampling technique, MapReduce, SVM.**

## I. Introduction

Imbalanced classification is the situation where the number of specimens with one class label is outstandingly lower than the specimens with other class label. In imbalanced dataset, the class with a relatively high number of specimens is called the majority class while the other one with less number of specimens is called the minority class [1][2]. This issue is prevalent in situations where an abnormal condition is vital, for eg., power pilferage, fake exchanges in banks, rare disease detection etc. In these situations, the traditional classification techniques might not give accurate results and in some cases the results can be predicted wrongly. This happens on the grounds that the traditional classification algorithms are normally intended to enhance accuracy by decreasing the error. This is because machine learning algorithms are designed to improve accuracy by reducing the error. Therefore, traditional classification models will not take imbalance ratio into consideration.

The class imbalance problem was handled at various stages like data level and algorithm level [22]. At the data level, solutions consist of different kinds of re-sampling techniques. At the algorithm level, we have cost learning algorithms which would incur more cost for misclassified minority (positive) samples and less cost for misclassified majority (negative) samples. Some of the Sampling approaches proposed are Random Over-Sampling [12][21], Random Under-Sampling [7][21] and SMOTE [4]. Popular cost learning approaches include SVM [17], k-Nearest Neighbour (kNN)[18], neural networks, genetic programming and rough set based algorithms. Under-sampling is better technique out of different traditional sampling techniques [28]. However, the existing imbalanced classification approaches cannot be adapted to the Big Data environment as the Map Reduce approach does not support these algorithms directly. Hence, a novel approach is proposed to handle the imbalanced classification problem on the massive amount of data by considering only essential features which would optimize the execution speed of the classifier and also improves its accuracy.

In this paper, the "Recursive Feature Elimination Under-Sampling" scheme is used to address the binary class imbalanced classification on the massive data. MapReduce framework with integration of statistical tool R is used for implementation. It is developed with a two-phase MapReduce. In the first phase, Feature Selection [2] is applied at every mapper to detect prominent attributes, then under-sampling is applied on the resultant dataset partition to balance it. On the resultant partitioned balanced dataset, SVM classification algorithm is applied to construct a model. Models generated by all mappers are saved for future use. In the second stage, saved models are tested with three UCI repository datasets [23]. UCI is the renowned Repository for real databases that are being used by various research communities for experimental analysis. It is observed that the accuracy of the new system is better when compared to a conventional SVM approach.

The rest of the work is organized as follows: Section 2 presents related work in imbalanced classification using sampling techniques. Section 3 describes Proposed architecture. Section 4 is about Experimental analysis and finally, Section 5 summarizes the conclusions and future scope.

## II. Related Work

This section reviews the Binary Imbalanced learning as well as MapReduce Framework.

E-ISSN 2581 – 7957
P-ISSN 2277 – 3916

CVR Journal of Science and Technology, Volume 17, December 2019
DOI: 10.32377/cvrjst1714

### *A. Binary Imbalance Classification Problem*

Binary imbalance classification is a two-class imbalanced learning problem where the total specimens in one class is extremely higher than the total specimens in other class.

### *B. Existing Solutions of Binary Imbalance Classification Problem:*

Adequate number of solutions have been proposed previously to handle the binary class imbalance classification issue. They are data pre-processing approach and algorithm approach. In data pre-processing approach, sampling technique is used to equalize the number of specimens in both the classes. Sampling techniques include Random Under-Sampling [7] [21], Random Over-Sampling [12][21] and SMOTE[4].

Sampling strategies focus on altering the distribution of training data either randomly, or by scaling existing specimens. This scaling can be done by either adding or removing existing specimens. Adding or removing of specimens depends on sampling technique used. In Random over-sampling, duplicate specimens are added randomly. In SMOTE over-sampling approach, k-Nearest neighbour specimens are added. In Random under-sampling, specimens that do not bring any improvement are removed.

#### *i) Random Oversampling*

Random Oversampling is a sampling technique where in a number of minority specimens are increased to balance the dataset. Random Over-Sampling [12][21] is simple to implement because minority specimens are increased just by duplicating existing specimens. However, it causes an overfitting because of the repetition of minority class instances. Random Over-Sampling is shown in the figure 1.

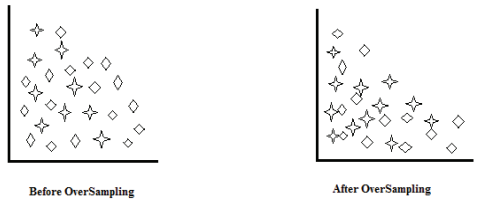


Figure 1.Random Oversampling

#### *ii)Random Undersampling*

It is the sampling technique wherein majority specimens are eliminated randomly [7] [21]. It randomly removes the majority class information to balance the dataset. However, it may discard the crucial specimens. In order to attack the issue of potential information loss, "near neighbor" method and its variations have been proposed. The basic algorithms of the near neighbor family are this: first, the method calculates the distances between all instances of the majority class and the instances of the minority class. Then k instances of the majority class that have the smallest distances to those in the minority class are selected. If there are n instances in the minority class, the "nearest" will result in k*n instances of the majority class.

"NearMiss-1" selects samples of the majority class that their average distances to three closest instances of the minority class are the smallest. "NearMiss-2" uses three farthest samples of the minority class. "NearMiss-3" selects a given number of the closest samples of the majority class for each sample of the minority class. Random Under-Sampling is shown in the figure 2.

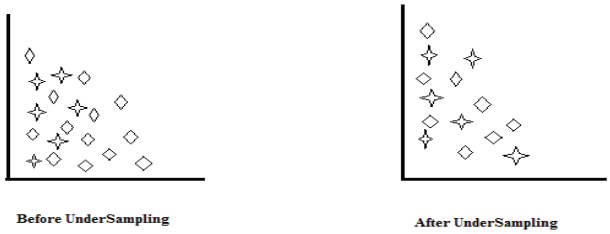


Figure 2.Random Undersampling

#### *iii) SMOTE: Synthetic Minority Oversampling*

SMOTE is an over-sampling approach proposed by Chawla et al [16],[4]. In SMOTE, *"synthetic"* specimens are created to increase the size of minority class. Unlike Random Over-Sampling that duplicates the specimens, SMOTE produce synthetic minority class specimens using k-Nearest neighbours, augmented with randomized interpolation. However, noise might be included in synthetic minority class examples.

### *C. MapReduce Framework*

The MapReduce framework [3] is used to process huge amounts of data in parallel. In Hadoop environment, data gets distributed among all nodes and it is processed in parallel.

The MapReduce framework [3] uses three phases like Map, Shuffle, Sort, and Reduce. Map is the first phase in MapReduce that divides input data into smaller and manageable sub-tasks to execute them in parallel. Then perform the required computation tasks. The output of the map is set of key, value pairs as <key, value>. Shuffle and Sort takes the output coming from Maps and perform the sub-steps on each (key, value) pair. It also returns <Key, List<Value>> output, but with sorted key-value pairs. Reduce is the final phase in the MapReduce framework. It takes a list of <Key, List<Value>> sorted pairs from shuffle function and performs reduce operation.

## III. PROPOSED SYSTEM ARCHITECTURE

In this section, architecture of the proposed system is designed as shown in Figure 3. Recursive Feature Elimination along with under-sampling is used to balance the data. The Support Vector Machine (SVM) is used for building binary classifiers.

E-ISSN 2581 – 7957
P-ISSN 2277 – 3916

CVR Journal of Science and Technology, Volume 17, December 2019
DOI: 10.32377/cvrjst1714

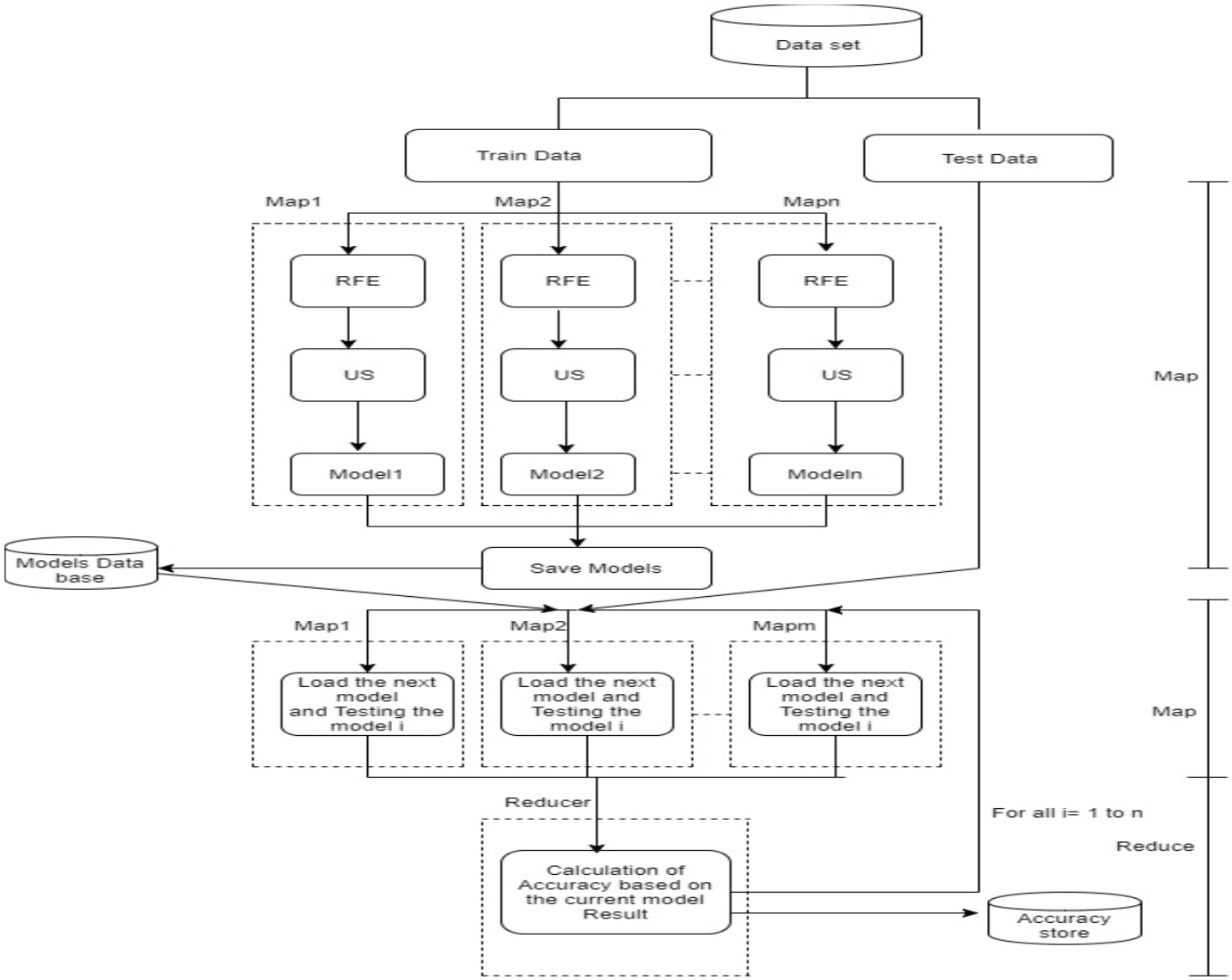


Figure 3..Proposed System Architecture

. The proposed system is developed in MapReduce paradigm. The Two-phase MapReduce scheme is used to develop the system. In the first phase of this scheme, training data splits among mappers, pre-processing is done at every mapper and then models are constructed and saved. In the second phase of MapReduce, the models are tested by dividing the test data among the mappers. Then the models are tested one by one iteratively. At the end of each iteration, reducer calculates model accuracy. The model which gives a better accuracy is considered as the resultant accuracy of the proposed system.

In the proposed system, models are saved for future use so that it supports "write once, use anywhere Principle" as long as the data to be tested is of the same domain.

### *A. Recursive Feature Elimination (RFE)*

Recursive Feature Elimination [19], recursively eliminates the non-essential features, the remaining attributes are used to build a model, and calculate the accuracy of the model. RFE can work on the combination of attributes that contribute to the prediction of the target variable. Finally, the subset based on error rate is obtained which is low. Algorithm for Recursive Feature Elimination is presented as below.

Algorithm 1: *Recursive Feature Elimination*

*Input: Dataset with a set of attributes*
*Output: Subset of the data of attributes*

1. *Define a list of features in order of dataset. like $f_1,f_2...f_n$*
2. *Resample the sample set by Random Forest, and method by cross validation.*
3. *For all the values of $f_i$ where i= 1 to n repeat*
   *3.i. Recursive Feature Selection for $f_i$*

E-ISSN 2581 – 7957
P-ISSN 2277 – 3916

CVR Journal of Science and Technology, Volume 17, December 2019
DOI: 10.32377/cvrjst1714

*3.ii. Calculate the error rate at each $f_i$ Level*

*4. Optimal feature number f*: the level with minimal cross validation error rate*

*5. Selected top features: the top feature f* highest-frequency feature.*

### *B. Under- sampling (US)*

The output of RFE would be the data with essential features. In under-examining, specimens of the lion's share class are chosen arbitrarily and the span of the greater part class is decreased closer to the measure of the minority class. The output of under-sampling is a balanced dataset. The yield of under-testing is adjusted dataset.

### *C. Support-Vector-Machine(SVM)*

Support Vector Machine (SVM)[17] is the supervised machine learning algorithm which can be used for both classification and regression problems. Kernel trick technique is followed in SVM to transform the data. The optimal boundary is found between possible outputs based on these transformations. In the proposed system, SVM is used to construct the model because it gives optimum classification result.

### *D. Model Save*

In the first phase of MapReduce, the model generated at the mapper is saved so that in the second phase of MapReduce this model is used for testing the data.

### *E. Model Load*

Models are tested in the second phase of MapReduce. For this, test data gets split among mappers, and then models are tested iteratively like the first model in the first iteration, the second model in the second iteration and so on. In each iteration, a model is tested in each mapper with the corresponding data partition. Output of each mapper is the confusion matrix of the corresponding data partition. A confusion matrix [27] is a table that is used to portray the performance of a classification model (or "classifier").

*Reducer Work*

In each iteration, output of mapper is given as the input to the Reducer. The Reducer does the calculations for accuracy by aggregating the output of each mapper. Whichever model gives an accurate result is considered to be the final accurate model of the proposed system.

## IV. Experimental Evaluation

In this section, the details of real-world problems having Binary-class imbalanced data are written, performance measures are explained and experimental study results are shown.

### *A. R-Hadoop integration*

The programming language used to implement the proposed system is statistical tool R [24] in integration with MapReduce, R is an open source programming dialect and gives programming condition to factual examination, designs portrayal and detailing. R is an open source tool that provides plenty of APIs to do statistical analysis, graphics representations.

### *B. Datasets and Parameters*

The datasets are taken from the UCI repository [23]. Table1 summarizes the details of selected datasets including a number of attributes and the Imbalance Ratio.

TABLE I.
Description About Datasets

| Datasets | Attributes Before RFE | Attributes After RFE | Imbalance Ratio |
|---|---|---|---|
| Breast Cancer Wisconsin (original)(Wobc) | 10 | 10 | 1.8 |
| Breast Cancer Wisconsin (Diagnostic)(Wdbc) | 32 | 16 | 1.6 |
| Breast Cancer Wisconsin (Prognostic)(Wpbc) | 34 | 20 | 3.2 |
| Page blocks0 | 10 | 10 | 8.79 |

### *C. Performance Measures*

Performance Measure evaluates how well an algorithm is performing on a given dataset. There are several performance measures exist in imbalanced classification like Precision/Specificity, Recall/Sensitivity, G-mean, F-measure, and AUC. However, for this study, accuracy is found using F-measure.

$$\text{Specificity} = TN/(TN+FP) \quad \text{.....(1)}$$

$$\text{Sensitivity} = TP/(TP+FN) \quad \text{.....(2)}$$

$$F\text{-measure} = 2X(\text{specificity } X \text{ sensitivity}) / (\text{specificity} + \text{sensitivity}) \quad \text{.....(3)}$$

### *D. Experimental Analysis*

We experiment on the datasets described in Table, UCI is the renowned Repository for real databases that are being used by various research communities for experimental analysis. The dataset of breast cancer samples are taken from the same repository. In this application 75% of the dataset is taken as training data to generate the models, and the rest 25% of the dataset is taken as the testing data.

The results of testing for all datasets described in Table I are summarized in Table II. From Table II, it can be analysed that the imbalanced classification accuracy is improved when SVM is combined with Under-Sampling and Recursive Feature Elimination. The improvements are highlighted.

CVR Journal of Science and Technology, Volume 17, December 2019
DOI: 10.32377/cvrjst1714

TABLE II.
THE RESULTS OF TESTING DATA

| Data Set | Accuracy of SVM | Accuracy of SVM + US | Accuracy of SVM + US + RFE |
|---|---|---|---|
| WOBC | 0.964271 | 0.960572 | **0.964271** |
| WDBC | 0.9785 | 0.9785 | **0.97995** |
| WPBC | 0.898305 | 0.898305 | **0.913792** |
| Page-blocks0 | 0.918760 | 0.918760 | **0.885168** |

It has been proved that the proposed algorithm works well to improve accuracy in case of Breast cancer dataset. At the same time, its performance is questioned for page blocks dataset. This is one possible area which has to be explored further and find out how the algorithm is actually working.

## V. CONCLUSIONS AND FUTURE WORK

In this paper, the parallelization scheme with Recursive Feature Elimination based on Under-Sampling is proposed for binary imbalanced classification using MapReduce. There is an improvement in the accuracy of the proposed system compared with the conventional SVM classification with Big Data. As part of future work, we would like to test the proposed system on various large datasets to study the consistency of the model.

## REFERENCES

[1] Nitesh V. Chawla, Nathalie Japkowicz, Aleksander Kolcz "Special Issue on Learning from Imbalanced Data Sets" Volume 6, Issue 1 - Page 1-6.
[2] Nitesh V. Chawla, Nathalie Japkowicz, Aleksander Kolcz ―Editorial: Special Issue on Learning from Imbalanced Data Sets‖ Sigkdd Explorations. Volume 6, Issue 1.
[3] A. H. The project, "Apache Hadoop," 2013. [Online]. Available: http://hadoop.apache.org/
[4] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," arXiv preprintarXiv:1106.1813, 2011.
[5] Triguero, D. Peralta, J. Bacardit, S. Garcia, and F. Herrera, "MRPR: MapReduce solution for prototype reduction in big data classification," Neurocomputing, vol. 150, pp. 331–345, 2015.
[6] Triguero et al., "Evolutionary undersampling for imbalanced big data classification," 2015 IEEE Congress on Evolutionary Computation (CEC), Sendai, 2015, pp. 715-722.doi: 10.1109/CEC.2015.7256961
[7] L. J. Eshelman, "The CHC adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination," in Foundations of Genetic Algorithms, G. J. E. Rawlins, Ed. San Francisco, CA: Morgan Kaufmann, 1991, pp. 265–283.
[8] Feature selection using Genetic algorithm is online: http://topepo.github.io/caret/feature-selection-using-genetic-algorithms.html
[9] W.-Y. Chen, Y. Song, H. Bai, C.-J. Lin, and E. Chang, "Parallel spectral clustering in distributed systems," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 33, no. 3, pp. 568–586, 2011. J. R. Cano, S. Garcia, and F. Herrera, "Subgroup discover in large size.
[10] T. Menzies, J. Greenwald, and A. Frank, "Data Mining Static Code Attributes to Learn Defect Predictors," IEEE Trans. Software Eng., vol. 33, no. 1, pp. 2-13, Jan. 2007.
[11] G. M. Weiss, "Mining with ararity: A unifying framework," SIGKDD Explorer, vol. 6, no. 1, pp. 7–19, 2004.
[12] J. Wang, M. Xu, H. Wang, and J. Zhang, ''Classification of imbalanced data by using the smote algorithm and locally linear embedding,'' in Proc. 8th Int. Conf. Signal Process., vol. 3. 2006, pp. 1–4
[13] C. S. Ertekin, ''Adaptive oversampling for imbalanced data classification,'' in Proc. 28th Int. Symp. Comput. Inf. Sci., vol. 264. Sep. 2013, pp. 261–269.
[14] H. He, Y. Bai, E. A. Garcia, and S. Li, ''ADASYN: Adaptive synthetic sampling approach for imbalanced learning,'' in Proc. IEEE Int. Joint Conf. Neural Netw., IEEE World Congr. Comput. Intell., Jun. 2008, pp. 1322–1328.
[15] I. Triguero, M. Galar, S. Vluymans, C. Cornelis, H. Bustince, F. Herrera and Y. Saeys, ''Evolutionary Undersampling for Imbalanced Big Data Classification,'' IEEE Trans.978-1-4799-7492-4/15/$31.00@ 2015 IEEE
[16] N.V. Chawla, N.Japkowicz, A.kolcz, Editorial:special issue on learning from imbalanced data sets, SIGKDD Ex[lor. Newsl.6(1)(2004)1-6.
[17] Ch.Sarada, M.Sathya Devi, "OVO Weighted Voting for Multi-Class Imbalanced Classification Having Distance as Weight", International Journal of Science & Engineering Research Volume 8, Issue 7,July-2017.
[18] KNN Classification is online at "http://www.math.le.ac.uk/people/ag153/homepage/KNN/OliverKNN_Talk.pdf".
[19] Xue-wen Chen,JongCheolJeong, "Enhanced recursive feature elimination", Machine Learning and Applications, 2007.
[20] Prediction matrix on online. "https://classeval.word press.com /introduction/basic-evaluation-measures/".
[21] G. Batista, R. Prati, and M. Monard, "A study of the behavior of several methods for balancing machine learning training data," ACM Sigkdd Explorations Newsletter, vol. 6, no. 1, pp. 20–29, 2004.
[22] Aida Ali, Siti MariyamShamsuddin and Anca L. Ralescu, "Classification with class imbalance problem: A Review" , Int. J. Advance Soft Compu. Appl, Vol. 7, No. 3, November 2015 ISSN2074-8523.
[23] UCI repository database at online, "https://archive.ics.uci.edu /ml/datasets.html".
[24] R description at online, "https://libguides.library.kent.edu/ statconsulting/r".
[25] R and Hadoop integration in online, "http://www.rdata-mining.com/big-data/r-hadoop-setup-guide".
[26] R Studio information at online, "http://dss.princeton.edu/train-ing/RStudio101.pdf".
[27] Confusion Matrix information in, "Simple guide to confusion matrix terminology", March 25, 2014
[28] Nadeem Qazi, Kamran Raza, "Effect Of Feature Selection, Synthetic Minority Over-sampling (SMOTE) And Under-sampling On Class imbalance Classification" 2012 14th International Conference on Modelling and Simulation, 978-0-7695-4682-7/12 $26.00 © 2012 IEEE