IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

Intrusion Detection on Highly Imbalanced Big Data using Tree Based Real Time Intrusion Detection System: Effects and Solutions

Dr.R.Balasubramanian[1], S.J.Sathish Aaron Joseph[2]

Research guide / Professor, P.G. and Research Department of Computer Science,

J.JCollege of Arts and Science (autonomous), Pudukkottai, Tamil Nadu, India [1]

Research Scholar / Head, Department of Computer Applications,

J.J.College of Arts and Science (Autonomous), Pudukkottai, Tamil Nadu, India [2]

Abstract: Due to the increased digitalization of information, a huge amount of data is being generated. Information richness in such data has attracted researchers to this data. The major problem existing in real time data is that it is usually huge and is imbalanced. This paper deals with analysing the tree based real time intrusion detection technique for intrusion detection from highly imbalanced Big Data. Classifiers tend to exhibit lower accuracies and reliabilities when the imbalance levels in the data are increased. Hence a highly imbalanced data is applied on the proposed classifier to determine its efficiency. Sampling techniques are some of the mostly used techniques to reduce the impact of imbalance on classifiers. Hence sampling techniques were applied on the data and the threshold limits for imbalance that can be effectively handled by the proposed classifier is identified.

Keywords:Classifier; Tree based Intrusion Detection; Sampling; Oversampling; Under Sampling; Imbalance; Big Data

I. INTRODUCTION

Cybersecurity has become a critical aspect due to the increase in the use of computers in all industries such as finances, medicine, industry etc. A major requirement in cyber security is Intrusion Detection. This acts as a major contributor in identifying attacks or malicious behaviour. Increase in efficiency of this system will lead to a better and more secure online environment. The ever increasing data plays a huge threat to this model. With the increase in usage of digital media, the data generated by them have also increased to a large extent. Hence faster processing with traditional approach is not feasible. It requires specific Big Data based approaches in order to effectively process them and provide results. A data is classified as Big Data if it effectively satisfies any one of the requirements of Big Data, namely; Volume, Velocity and Variety. Real time data generated satisfies the constraint of Big Data, hence it is classified as Big Data.

Real time data in specific exhibit several practical issues. One major issue is the presence of imbalance especially in applications like intrusion detection. A data set is said to be imbalanced if one of its classes plays a huge dominance over the other existing classes. The class that dominates the dataset is called the majority class, while the other classes are called minority classes. The unequal distributions can be in any level, ranging from 1:2, 1:100, 1:1000, 1:10000 etc. It was observed from several real time datasets that this ratio can even be very huge in terms of 1:100000 [1]. Some real time applications where such scenarios exist includes anomaly detection, classification of documents [2,3,4], gene classifications, image classification [5], identification of frauds in banking or telephone calls [6], biomedical applications to identify rare genetic disease etc. The applications listed here clearly depict the levels of imbalance that can be expected of in real time data sets. Further, the major problem here is that the minority classes are of higher importance, while the majority classes occupy the least importance. Hence identifying the minority classes effectively should be the major concern of any classifier, since the cost of misclassification of minority classes is higher than that of the majority classes [7].

But such imbalance levels tend to affect classifiers by biasing them towards the majority classes. This paper discusses the effects of imbalance on the enhanced tree based real time intrusion detection system by applying data sets of varying imbalance. Methods to counter imbalance are also discussed in detail and sampling techniques were identified to be the best methods to counter imbalance. The data set is then sampled to various levels and effects of sampling on the classifier were analyzed. Threshold choke points were identified and best sampling levels were identified.

II. INTRUSION DETECTION ON HIGHLY IMBALANCED BIG DATA USING TREE BASED REAL TIME INTRUSION DETECTION SYSTEM: EFFECTS AND SOLUTIONS

Real time intrusion detection tends to be a tedious task due to the large amount of data involved. Though this could be solved by improving the computational capabilities of the systems, another major hurdle tends to be data imbalance. This section presents the effects data imbalance on the tree based real time classifier proposed by the authors in [8].

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

A. Enhanced Tree Based Real Time Intrusion Detection System

The enhanced tree based intrusion detection system [8] uses an ensemble of decision trees to perform classification. Size of the ensemble classifier is determined and the training data is split accordingly, such that every decision tree in the classifier is provided with at least 66% of the training data. The reason for such division is that every class should contain their representations in each decision tree of the ensemble classifier in order to perform efficient classification. Each decision tree identifies a subset of m predictor variables from a list of M total predictors. The best predictor variable is identified from the set of m values and a split is performed on it. This process is repeated for all the predictor variables and a decision tree is constructed. Pruning is avoided in order to reduce information loss. Every decision tree operates on the data provided to them, hence if the ensemble contains k independent decision trees, k different rules are finally generated. Since a subset of data was used for training, none of the decision rules obtained at this stage are complete. These rules obtained at the intermediate stage are referred to as weak rules. All the generated rules are aggregated to obtain the final decision tree, called the strong classifier. This method works on the basic principle that several weak classifiers can be combined to form a strong classifier.

B. Effects of Imbalance on Tree Based Real Time Intrusion Detection System

Imbalance is the inevitable problem occurring in real time data due to the huge size and low frequency of certain transactions. This paper deals with intrusions, which are usually anomalies whose frequency of occurrence is very rare. Hence our approach of tree based real time IDS has very high probability of being impacted by imbalance. Results from [8] indicated that low to moderate imbalance levels did not affect the accuracy or the reliability of the classifier. This can be attributed to the fact that several classifiers operate on parts of data, hence even if one classifier has low representations of a data, the other classifiers would contain sufficient representations. Hence when the results are combined to obtain the strong classifier, effects of imbalance are either reduced to a large extent or most probably neutralized.

Figures 1 and 2 shows the accuracy levels and the reliability levels obtained on datasets with moderate to high imbalance. KDD Cup 99 is used as a representative set for high imbalance. The imbalance level observed in KDD Cup 99 was observed to be 3274:1. Hence for every attack record, there exists 3274 normal records. High accuracy levels of 99.9% are exhibited by the classifier even with the huge imbalance levels. This is completely attributed to the fact that due to the high imbalance levels, even the test set has low representations of the minor class. Table 1 presents the confusion matrix corresponding to Figure 1.

Prediction results observed from Table 1 shows that all the records were predicted to be positive or in other words all the records were predicted to be a part of the major class. All entries corresponding to the minor class was wrongly predicted. This is clearly exhibited from Figure 2, which shows the reliability levels of the classifier. The classifier for KDD Cup 99 dataset exhibited a reliability level of 33.3%, which is the lowest among all the representations.

Table 1: Confusion Matrix

Data Set	TP	FP	TN	FN
Glass1	9	2	19	2
Yeast1	30	13	145	34
Vehicle0	21	1	85	1
Ecoli	3	2	36	1
Glass5	1	0	30	1
Yeast5	4	2	181	2
KDD	145914	42	0	0

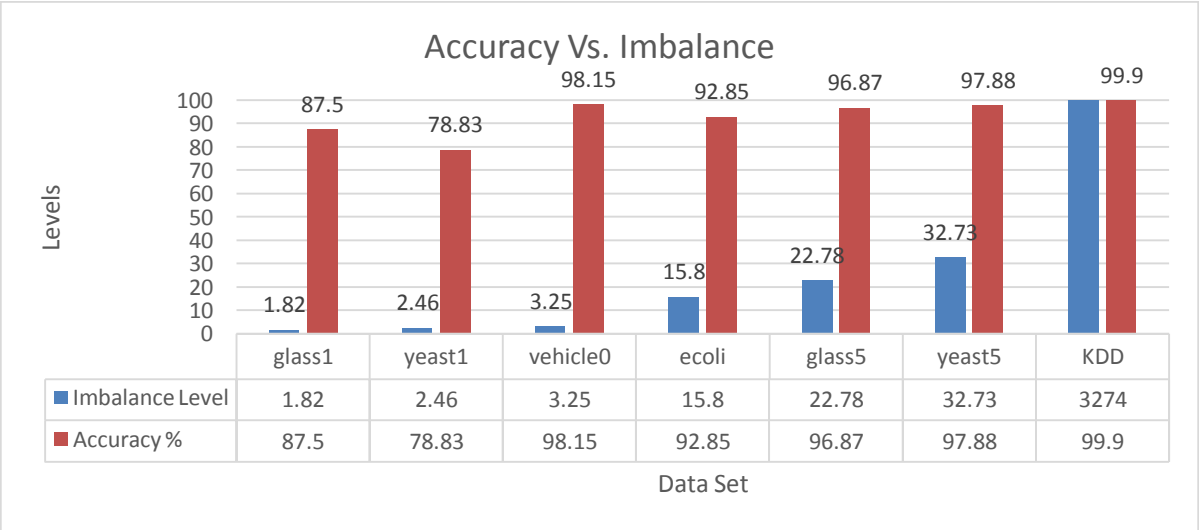

Fig 1. Accuracy Vs. Imbalance

DOI 10.17148/IJARCCE.2016.5208

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

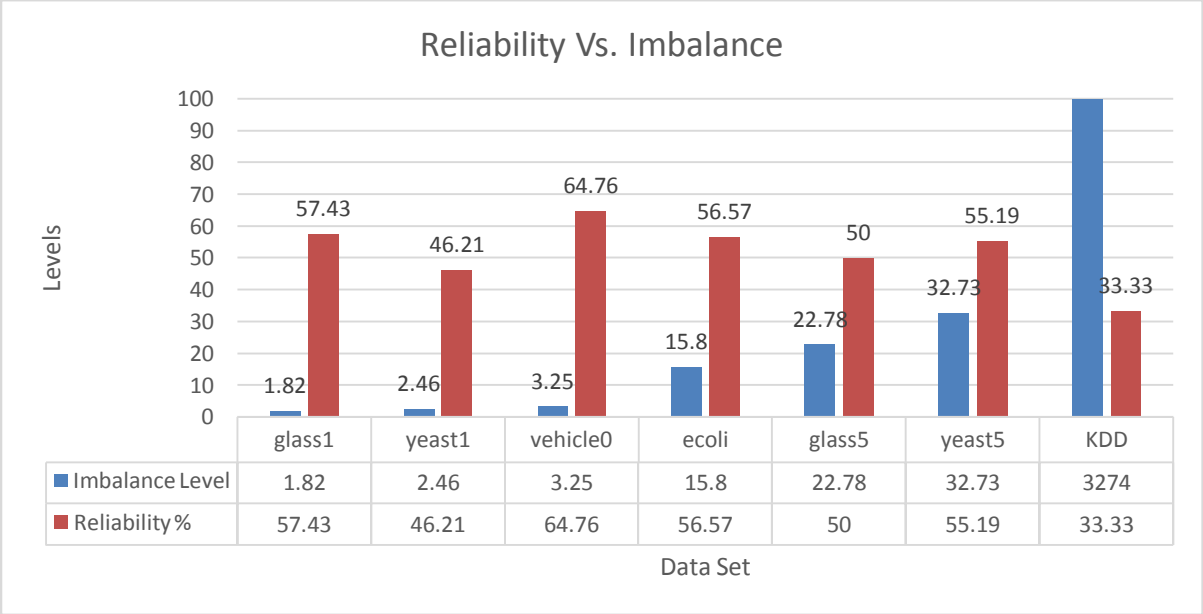

Fig 2. Reliability Vs. Imbalance

Though it might be argued that utilizing more than 66% of the data for each decision tree would increase the accuracy, it should also be considered that higher the percent of data provided to the decision trees, more similar decision rule would be generated. This will downgrade the entire purpose of utilizing the ensemble technique. Providing the entire data to all the classifiers is also not feasible, since the application is dealing with Big Data. Hence reducing the level of imbalance is the only feasible solution to improve the reliability of the classifier.

C. Imbalance: Solutions

Countering imbalance is one of the major areas of research, where real time applications are involved. The major problem to be faced is that, by default the classifiers consider the data provided to them to be balanced. This behavior biases the classifiers towards the majority classes, while the minority classes are being neglected. This behavior was clearly visible from the Confusion matrix in Table 1. Two major solutions exist for this problem; creating a weighted classifier such that it provides more importance to the minority classes [9] or applying sampling techniques on the data to balance the dataset [2, 10, 9, 11]. This section discusses the sampling techniques available and their operational schemes.

I. Oversampling

Oversampling is the process of increasing the size of the minority classes such that it balances with the majority classes. These techniques increase the amount of data by creating artificial data points with the already available data, such that the data consistency is maintained. A widely used oversampling technique is SMOTE [12] proposed by Nitesh et al. This method uses pairs of minority class samples to generate the artificial data point such that it falls between the selected samples, hence creating no impact in the consistency of dataset. Several oversampling techniques proposed in literature for handling data imbalance includes [13-15].

II. Under sampling

Under sampling is the converse of oversampling, where the majority classes are eliminated to equalize the imbalance ratio. The effectiveness of undersampling has been analyzed and presented in [16]. This method has its major concentration on the process Classification. A hybrid classification approach that utilizes both the weighing scheme and the undersampling technique is presented in [17]. Some recent contributions to undersampling techniques also include utilizing metaheuristics to perform undersampling [18], rather than the legacy methods of utilizing random elimination or elimination using statistical analysis.

Apart from the above mentioned categories, a combination of oversampling and under sampling, called the hybrid sampling technique [11, 19] is also on the raise. These methods tend to utilize the advantages of both the sampling techniques in order to compensate for the downsides.

D. Effects of Sampling on Accuracy and Reliability

This section presents the impact of sampling techniques on the accuracy and the reliability of the classifier. Oversampling and under sampling techniques were applied on the KDD Cup 99 dataset. Datasets with imbalance levels of 3000, 2000, 1000, 100, 50, 10 and 1 were obtained from both oversampling and under sampling and the results were recorded. SMOTE was used for oversampling, while random under sampling technique was used for under sampling the results.

Fig 3 presents the accuracy and reliability levels of the sampled dataset. The center point refers to the actual KDD dataset, representing 99.9% accuracy and 33.3% reliability. The data represented above the KDD dataset depicts results from under sampled data, while the data represented below the KDD dataset represents results from oversampled data.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

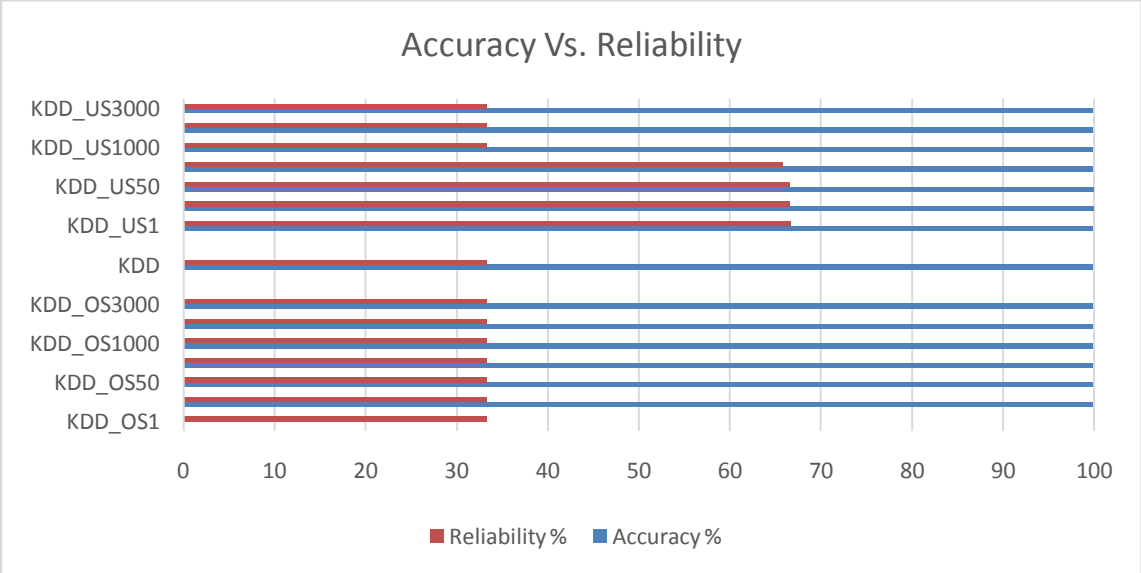

Fig 3. Accuracy Vs. Reliability

Table 2: Accuracy and Reliability of Sampled Data

DataSet	TP	FP	TN	FN	Accuracy %	Reliability %
KDD_OS1	0	0	42	145914	0.0288	33.33
KDD_OS10	145914	42	0	0	99.9	33.33
KDD_OS50	145914	42	0	0	99.9	33.33
KDD_OS100	145914	42	0	0	99.9	33.33
KDD_OS1000	145914	42	0	0	99.9	33.33
KDD_OS2000	145914	42	0	0	99.9	33.33
KDD_OS3000	145914	42	0	0	99.9	33.33
KDD	145914	42	0	0	99.9	33.33
KDD_US1	145870	0	42	44	99.9	66.66
KDD_US10	145914	0	42	0	100	66.6
KDD_US50	145914	0	42	0	100	66.6
KDD_US100	145914	1	41	0	99.9	65.87
KDD_US1000	145914	42	0	0	99.9	33.3
KDD_US2000	145914	42	0	0	99.9	33.3
KDD_US3000	145914	42	0	0	99.9	33.3

It was observed from the oversampled results that the accuracy and reliability remained the same as with the KDD data. It could also be observed that the oversampled dataset exhibiting an imbalance level of 1 exhibits an accuracy of 0.0288%, which is attributed to the hugeness of the data (Table 2).

Under sampling exhibits much promise. Under sampling datasets with ratios 3000, 2000 and 1000 exhibits accuracy and reliability similar to the actual KDD data. However as the imbalance level reaches 100 and less, the reliability levels exhibits an increase of upto 66%. This depicts the throttle point of imbalance.

Figures 4 and 5 represent the accuracy and reliability graphs in terms of the imbalance levels.

It could be observed from Figure 3 that an increase in reliability can be observed after an imbalance level of 100 (Under sampled), hence defining the maximum threshold limit of imbalance that can be handled by the tree based real time intrusion detection system.

The accuracy graph (Figure 4) shows a stable accuracy of 99.9% on all the datasets except for the oversampled KDD dataset with imbalance 1, depicting errors occurring due to overtraining.

DOI 10.17148/IJARCCE.2016.5208

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

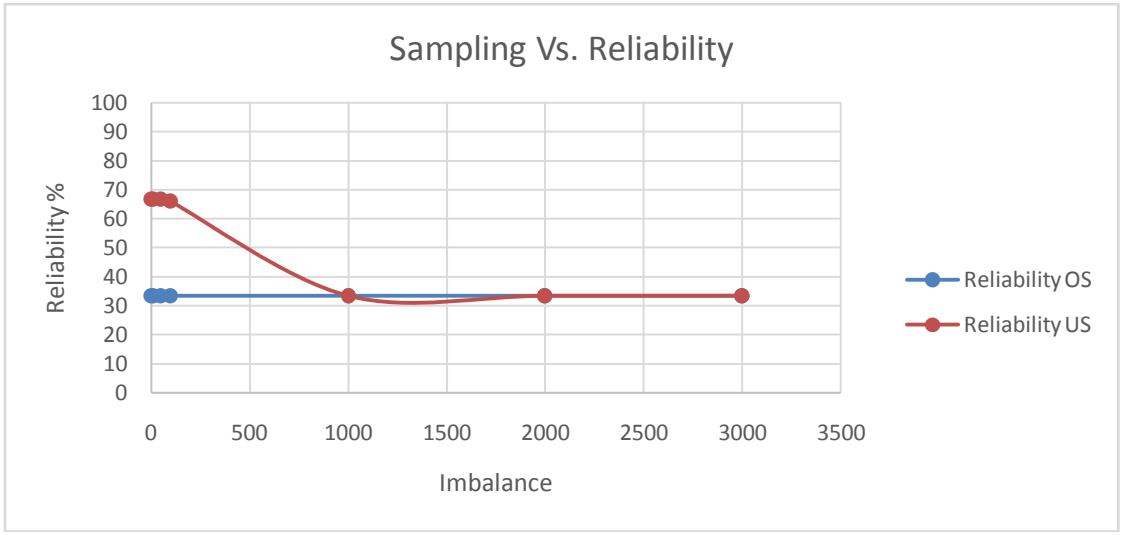

Fig 4. Sampling Vs. Reliability

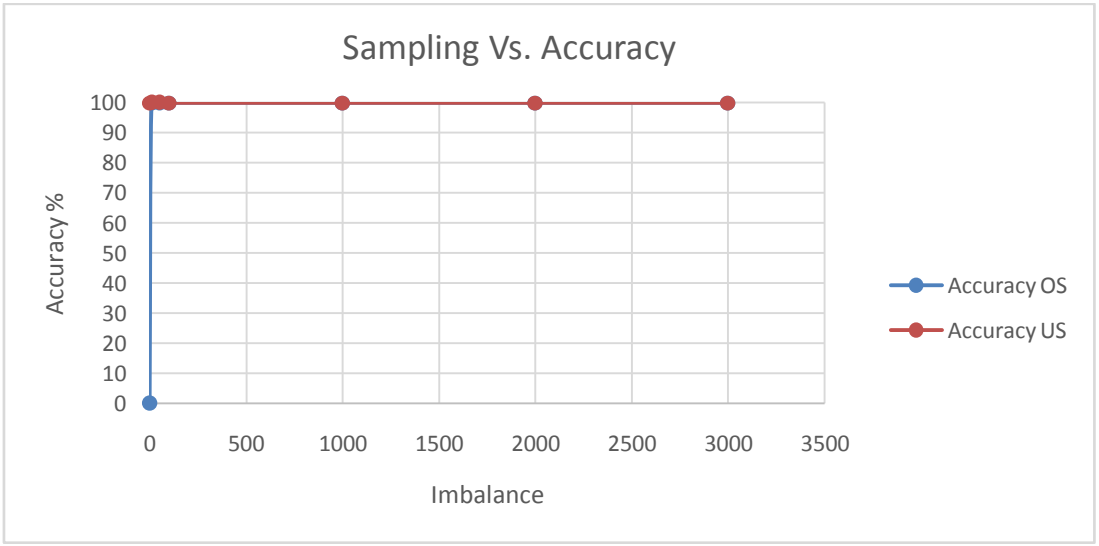

Fig 5. Sampling Vs. Accuracy

III. CONCLUSION

This paper presents an analysis of the effects of imbalance on the enhanced tree based real time intrusion detection system. The proposed system, being a real time intrusion detection system will certainly contain imbalanced data, hence this analysis becomes mandatory. Imbalance and its effects on the classification accuracy and reliability have been discussed in detail.

An effective solution to handle imbalance; sampling was also discussed in detail. Types of sampling and the impact sampling has on the accuracy and reliability of the classifier has been experimentally evaluated. Threshold limits of imbalance handles by the tree based intrusion detection system has been identified. Future direction of research includes proposing algorithms for pushing the threshold limits such that imbalance of higher levels can be handled effectively. The current approach has been specifically developed for binary classifiers. Future enhancements can also be made by extending this approach to support multiclass classifiers.

REFERENCES

[1] F. Provost, and T. Fawcett,"Robust Classification for Imprecise Environments", Machine Learning, 42/3, 203–231, 2001.

[2] D. Lewis, and J. Catlett, "Heterogeneous Uncertainity Sampling for Supervised Learning", In Proceedings of the Eleventh International Conference of Machine Learning, pp.148–156 San Francisco, CA. Morgan Kaufmann, 1994.

[3] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive Learning Algorithms and Representations for Text Categorization". In Proceedings of the Seventh International Conference on Information and Knowledge Management, pp. 148–155, 1998.

[4] D Mladeni´c, and M. Grobelnik, "Feature Selection for Unbalanced Class Distribution and Naive Bayes". In Proceedings of the 16th International Conference on Machine Learning, pp. 258–267. Morgan Kaufmann, 1999.

[5] K, Woods, C, Doss, K. Bowyer,J. Solka, C. Priebe, and P. Kegelmeyer, "Comparative Evaluation of Pattern Recognition Techniques for Detection of Microcalcifications in mammography", International Journal of Pattern Recognition and Artificial Intelligence, 7(6), 1417–1436, 1993.

[6] T. Fawcett, and F. Provost, "Combining Data Mining and Machine Learning for Effective User Profile", In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, pp. 8–13 Portland, OR. AAAI, 1996.

DOI 10.17148/IJARCCE.2016.5208

[7] N. Tomašev, and D. Mladenić, "Class imbalance and the curse of minority hubs." Knowledge-Based Systems 53 (2013): 157-172.
[8] S.J.Sathish Aaron Joseph, R. Balasubramanian, "Enhanced Tree Based Real Time Intrusion Detection System in Big Data", International Journal of Computers & Technology, 2010.
[9] N. Japkowicz, "The Class Imbalance Problem: Significance and Strategies", In Proceedings of the International Conference on Artificial Intelligence : Special Track on Inductive Learning Las Vegas, Nevada, 2000.
[10] M. Kubat, and S. Matwin, "Addressing the Curse of Imbalanced Training Sets: One Sided Selection". In Proceedings of the Fourteenth International Conference on Machine Learning, pp. 179–186 Nashville, Tennesse. Morgan Kaufmann, 1997.
[11] C. Ling, and C. Li, "Data Mining for Direct Marketing Problems and Solutions", In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98) New York, NY. AAAI Press, 1998.
[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique", Journal of artificial intelligence research, 321-357, 2002.
[13] I. Nekooeimehr, and S. K. Lai-Yuen. "Adaptive semi-unsupervised weighted oversampling (A-
[14] SUWO) for imbalanced datasets." Expert Systems with Applications 46 : 405-416,2016.
[15] H. Zhang, and L. Mingfang, "RWO-Sampling: A random walk over-sampling approach to imbalanced data classification." Information Fusion 20: 99-116,2014.
[16] M. Gao, . "PDFOS: PDF estimation based over-sampling for imbalanced two-class problems." Neurocomputing138 : 248-259,2014.
[17] A. Dal Pozzolo, O. Caelen, and G. Bontempi. "When is undersampling effective in unbalanced classification tasks?." Machine Learning and Knowledge Discovery in Databases. Springer International Publishing 200-215, 2015
[18] A. Anand, et al. "An approach for classification of highly imbalanced data using weighting and undersampling." Amino acids 39.5 : 1385-1391,2010.
[19] F. Charte, et al. "MLeNN: a first approach to heuristic multilabelundersampling." Intelligent Data Engineering and Automated Learning–IDEAL, Springer International Publishing, 1-9.2014.
[20] S. Schistad, H. Anne, and S. Rune, "A large-scale evaluation of features for automatic detection of oil spills in ERS SAR images." Geoscience and Remote Sensing Symposium, IGARSS'96.'Remote Sensing for a Sustainable Future.', International. Vol. 3. IEEE, 1996.

BIOGRAPHY

The author **S.J.SATHISH AARON JOSEPH** is working as the head of the department of computer Applications in J.J.College of Arts and Science (Autonomous), Pudukkottai, Tamil Nadu, India. At present he is a part time research scholar in Ph.D computer science under the guidance of **Dr.R.Balasubramanian**, Professor, P.G and Research Department of Computer Science, J.J.College of Arts and Science (Autonomous), pudukkottai,Tamil Nadu.

DOI 10.17148/IJARCCE.2016.5208