Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data

Conference Paper · January 2016

CITATIONS
14

READS
3,421

2 authors:

Akila Somasundaram
National Institute of Technology Tiruchirappalli
7 PUBLICATIONS **56** CITATIONS
SEE PROFILE

U. Srinivasulu Reddy
National Institute of Technology Tiruchirappalli
33 PUBLICATIONS **110** CITATIONS
SEE PROFILE

Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data

Akila Somasundaram
Research Scholar,
Computer Applications, National Institute of Technology,
Trichy, India
akila29@gmail.com

U. Srinivasulu Reddy,
Assistant Professor,
Computer Applications, National Institute of Technology,
Trichy, India

***Abstract*—Big Data and Big Data Analytics has gained huge prominence over the recent years. Their ability to handle massive amounts of data at ease has made them the preferred choice for many real time applications. Some major problems existing in normal data such as data insufficiency are implicitly absent in Big Data. This paper focuses on one such problem existing most prominently in real life applications, data imbalance. This problem arises due to the skewed nature of certain real time applications. The proposed approach analyses data imbalance and its impact on the process of classification. Map Reduce implementations of Naïve Bayes Classifier and Random Forest Classifier were used on Big Data to identify accuracy and reliability exhibited by them when the training provided to the classifier model is biased. Data sampled with several levels of imbalance were used for analysis. Experiments were conducted and the results exhibited were analyzed to identify the threshold levels of imbalance, best sampling technique for Big Data and the comportment of algorithms towards imbalanced data.**

Keywords— Big Data; Classification; Data Imbalance; Fraud Detection; Sampling

I. INTRODUCTION

Big Data being the current buzzword presents itself as the best solution for problems brought forward by the hugeness of the data. Support for Volume, Velocity and Variety in data has made Big Data Technologies the most eligible candidates for data analytics [1]. This machine generated data can be financial transactions, data from social networking media, gene sequences, sensor generated data, geographical/ geological image data, webcam generated images for face recognition etc. which requires the analytics technique to identify transactions of importance from regular patterns. Interesting patterns are to be extracted from huge data sets for analysis. These interesting patterns tend to occupy < 1% of the data in most real applications, which complicates the identification process. Identifying the occurrence of these interesting patterns in the currently generated data usually performed in a supervised manner. Classification is the supervised learning technique used to identify the category to which a new observation belongs to, from a set of defined categories. The areas in which classification techniques are utilized include anomaly detection, document classification [2, 3, 4], image classification [5], gene classification etc.

Sparse availability of interesting patterns in the data leads to data imbalance. Though sampling techniques are available to counter imbalance, the question arises as to what level of sampling is tolerated by the Classification techniques. The effects of imbalance on big data and an analysis on the sampling levels tolerated by classification techniques operating on big data are discussed in this paper. This paper approaches data imbalance from the perspective of Big Data and analyses the effects of sampling on imbalanced data when Big Data Analytics are applied on them. This paper is the initial phase of analysis for the work of non- intrusive credit/debit card fraud detection and prevention.

The remainder of this paper is structured as follows; Section II presents the effects of imbalance on the process of classification, Section III provides methods to counter imbalance, Section IV presents the experimental results, Section V presents the discussion and Section VI concludes the study.

II. DATA IMBALANCE AND ITS EFFECTS ON CLASSIFICATION

A. Data Imbalance

Data imbalance is one of the major problems prevailing in real time anomaly detection datasets. A dataset is considered to be imbalanced if one of its classes plays a huge dominance over the rest of the classes. This property can be observed with most prominence in binary classification datasets, as most Binary classification datasets are implicitly imbalanced. The major class entries are higher, while the minor class entries occupy a very small space in the dataset. This unequal distribution can be of the form 100:1, 1000:1, 10000:1 etc. Several real time data sets [6] has shown even higher imbalance ratios of 100000 to 1. The areas in which such imbalance levels can be observed include fraudulent telephone calls or transactions [7], banking frauds, bio-medical applications for identifying rare genetic disease, document filtering, remote sensing, defect in computer industry and other manufacturing industries. From the applications, it could be observed that the rare or minor occurrences are of higher importance, while the major occurrences represent normal observations.

ISBN NO: 978-81-908388-7-0

Imbalanced data are observed to follow certain characteristics namely

a. Overlapping small disjuncts

b. Lack of density

c. Noise

d. Dataset shift [20]

These characteristics of imbalance data make the process of classification on such data a difficult task.

B. Effects of Data Imbalance on Classification

Classification, being a supervised learning process depends mainly on the training data. The level of training plays a major role in the resultant accuracy of the classifier. Imbalanced nature of the data sets acts as a huge downside in this scenario. Due to the minimal occurrence of the minor classes, the classifier gets insufficiently trained and hence provide inaccurate predictions. In case of multi class classifiers, such imbalance results in low representations from entries and eventually these entries tend to get totally ignored [23]. Most classifiers tend to implicitly consider their data as balanced, hence standard classifiers are biased towards the majority.

The effects of this problem are observed to be minimal when observed from the perspective of accuracy of the classifier. Occurrences of false positives or false negatives is only considered as a part in determining the accuracy, hence its impact is reduced. Due to the imbalanced nature of the data, the minimal entries depicting minor classes reduces this further. When considering a dataset with an imbalance ratio 100:1, it is sufficient to label all the entries as positive to obtain 99% accuracy from the classifier. As the imbalance level increases, the accuracy is bound to increase. Though this level of accuracy is acceptable from a theoretical perspective, in the real world, the cost of misclassification is high especially in the generation of false positives. This leads to a scenario where the major focus is to be placed on the minority class during the training phase and the application tolerates a small error rate in the majority class [24].

Data relating to Credit/ Debit card transactions tend to contain huge imbalance levels. This complicates the process of fraud detection. It was also observed that while high accuracy levels were observed, the reliability levels remained low. A sample of 1 Million records were considered and the accuracy and reliability levels were plotted in Figure 1. The classifier accuracy remained stable (99%) and the imbalance levels were increased from 1, 5, 10, 50, and 100 until 999999.

The algorithm exhibited tolerance towards imbalance until an imbalance level of 10, after which the reliability levels started dropping, and reached a level of 0.1, when the imbalance level is moved to 500. On further increase, the reliability levels dropped down to 0.0001 (for an imbalance ratio of 1 Million). The major downside is that most of the real life applications have such imbalance levels [25]. This reveals that the accuracy of a classifier is not a metric to be completely trusted especially when it comes to data with imbalance.

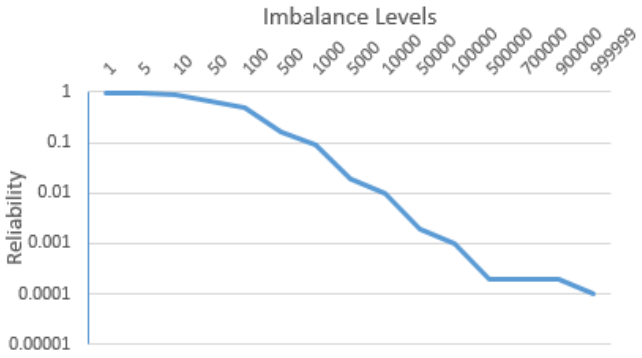

Fig. 1. Effect of Imbalance on Reliability

III. TECHNIQUES TO COUNTER IMBALANCE

Countering imbalance in data tends to be one of the major areas of research when real time classification is concerned. Classifiers operating on data have a basic assumption that the data is balanced. Hence the weight provided to each of the samples is equal [26]. However, in imbalanced data, this mode of operation makes the classifier biased towards the majority classes. Provided with a sufficiently high level of imbalance, the minority classes can even get ignored during the rule building process. Data balancing techniques [21] have been proposed to handle this scenario. Data handling can be performed either by modifying the existing algorithms to provide increased weightage to the minority classes [8], making their contribution levels higher, or by sampling [2, 9, 10, 11]. This paper deals with the analysis of the second category of modifying the existing dataset to counter the imbalance.

A. Oversampling

Oversampling refers to increasing the size of the minority class to balance the majority class. This method tends to duplicate the data already available or generate data on the basis of available data. SMOTE [14] one of the popular oversampling technique was proposed by Nitesh et al. This method uses pairs of minority class samples to generate data satisfying the conditions for minority classes. It operates on the feature space, rather than the data space. The synthetic samples that are created causes classifiers to create larger and less specific decision regions, rather than smaller and more specific regions. SMOTE was inspired from an oversampling technique proposed for handwriting recognition [15].

B. Under-sampling

Under-sampling reduces the size of the majority classes to balance them with the minority ones. The level of data loss depends mainly on the measure used for eliminating records. Geometric mean based under-sampling technique was used by Kubat and Matwin [9]. Another related work called the SHRINK system was proposed by Kubat et al., in [12] for the detection of oil spills.

C. Hybrid Sampling

Several recent contributions have come up with hybrid sampling techniques that combine oversampling and under-sampling to provide a balanced dataset. A combination of oversampling and under-sampling techniques are used in the recent studies citing a clear advantage over the utilization of any one of these techniques. One such approach is that of Ling and Li proposed in [11]. Lift curve is used as the analysis criteria in this method. Another hybrid method for oil slick classification from SAR imagery is proposed by Solberg [13].

Identifying the effect of imbalance in a dataset on the basis of under-sampling, resampling and recognition based induction scheme was proposed by Japkowicz [10]. This method applied random resampling, focused resampling and random under-sampling on artificially generated 1D data.

D. Results

Experiments were conducted with the Credit Card Fraud dataset from Brazilian Bank [17] and the KDD CUP 99 dataset (partial). The bank dataset contains 345,719 data records, with an imbalance ratio of 25. Since the current paper deals with binary classification problem, two classes (normal and smurf attack) were selected from the KDD CUP 99 dataset. The KDD CUP dataset contains 827,089 entries with an imbalance level of 3274. Oversampling and under sampling were applied on them to reduce the imbalance levels. ROC and PR curves were used to measure their performance.

Experiments were conducted on a 12 node Hadoop cluster (10 Data Nodes + 1 Name Node + 1 Secondary Name Node). Map Reduce implementation of Naïve Bayes was used on the data to observe the effects of imbalanced data on a statistical classifier. Map Reduce and Spark based implementation of Random Forest [15] from Mahout 0.9 was used to identify the effects of imbalance on a Random Forest classifier.

ROC [19] and PR curves were used to visualize the performance of the classifiers. PR curves are used to uncover information that are not available from ROC curves [18]. They work well on skewed datasets. Due to the high level of imbalanced data handled by this approach, it becomes mandatory to examine the PR space [22].

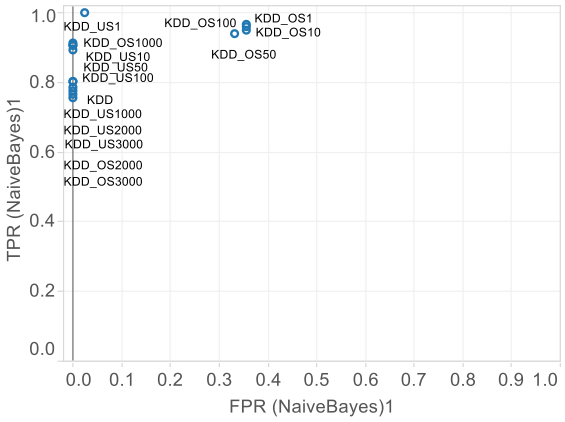

Fig. 2. ROC Curve for Naïve Bayes Classifier on KDD CUP 99 data

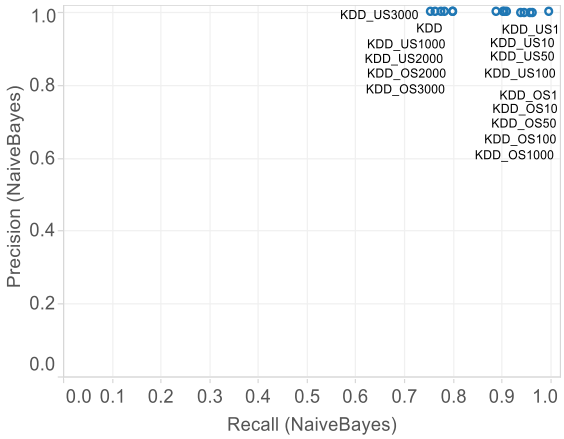

Fig. 3. PR Curve for Naïve Bayes Classifier on KDD CUP 99 data

The plots in Figures 2 and 3 shows the ROC and PR plots for the Naïve Bayes Classifier (NBC) on the actual KDD data, the oversampled and undersampled KDD datasets with imbalance ratios ranging from 3000 to 1 (3274 is the actual imbalance).

It could be observed that in the ROC plot KDD oversampled to an imbalance level 1 exhibits a False Positive Rate (FPR) of 0 and True Positive Rate (TPR) of 0.77. The PR plot exhibits high precision and above average recall, of 0.77. On analysing the oversampled results, it was observed that they exhibited high true positive rates (0.9), while the lowest of 0.8 and 0.7 was exhibited by oversampled records with imbalance of 2000 and 3000. The PR curves also exhibits similar results. On examining the undersampled results it was observed that the dataset exhibiting an imbalance ratio of 1 exhibits best performance with low FPR (0.02), high TPR (0.99), high precision (0.99) and recall (0.99). it exhibits an incorrect classification rate of 0.3%. All the other datasets, though they exhibit low FPR (0) and high Precision(1), they exhibit moderate TPR (0.75 to 0.91).

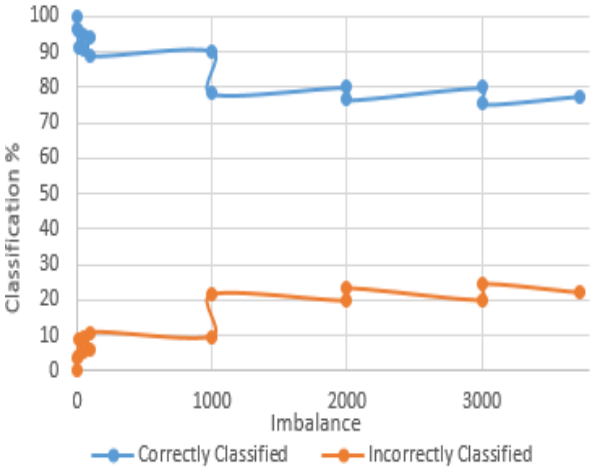

Fig. 4. Imbalance (1-3724) Vs. Classification Accuracy (NBC)

ISBN NO: 978-81-908388-7-0

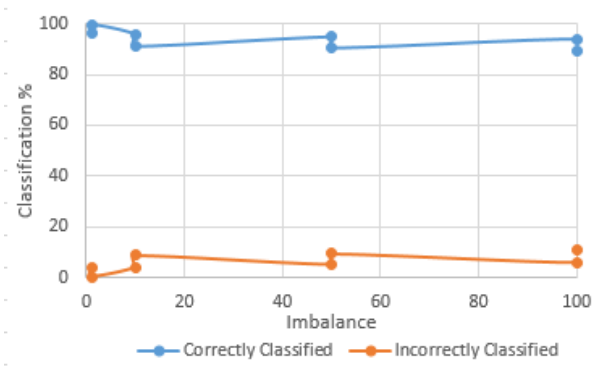

Fig. 5. Imbalance (1-100) Vs. Classification Accuracy (NBC)

Classification accuracies with respect to imbalance levels is presented in Figure 4. It was observed that the low levels of imbalance (1-100) has less impact on the accuracy levels. As the imbalance levels increase, variations observed were linear. Figure 5 shows a cropped view of the data representing data with imbalance levels from 1-100. The overall transition in this interval was observed to be very low. True Negative Rate (TNR) and the Negative Predictive Value (NPV) exhibit the actual impact on the minor classes. Oversampled results (1-100) exhibited very low TNR, while others exhibited high TNR. NPV remained very low (0.001 – 0.08) and Mathews Correlation Co-efficient (MCC) also remained low approximating to 0, indicating random prediction.

Figure 6 shows the aggregated ROC points obtained from oversampled and under sampled KDD datasets using Random Forest as the classifier. Every point in the plot corresponds to a single dataset. Other results obtained from the classifier are tabulated in Table 2. It was observed that in terms of ROC, best results were exhibited by datasets with imbalance ratios of 1, 10, 50 and 100. They exhibit almost zero false positive rates, which is the ideal scenario. Other datasets exhibit high false positives and true positives. The worst scenario was observed in the oversampled set with imbalance of 1.

Table 1: Results for Naïve Bayes Classifier

Data Set	Accuracy %	Kappa	True Negative Rate (TNR)	Negative Predictive value	Mathews Correlation coefficient
KDD	77.58	0.002	1	0.001281582	0.031529509
KDD_OS1	96.31	0.0094	0.642857	0.005000926	0.054461824
KDD_OS10	95.85	0.0083	0.642857	0.004449572	0.051105072
KDD_OS50	94.75	0.0064	0.642857	0.003516083	0.044854231
KDD_OS100	93.93	0.0057	0.666667	0.003155285	0.043015588
KDD_OS1000	90.32	0.0053	1	0.002964008	0.051739935
KDD_OS2000	80.04	0.0023	1	0.001439984	0.033949334
KDD_OS3000	79.93	0.0023	1	0.001432176	0.03383412
KDD_US1	99.69	0.1537	0.97619	0.083673469	0.28533678
KDD_US10	91.14	0.0059	1	0.003237493	0.054319282
KDD_US50	90.5	0.0055	1	0.003020931	0.052287279
KDD_US100	89	0.0046	1	0.002610155	0.048198187
KDD_US1000	78.27	0.0021	1	0.001322418	0.032170763
KDD_US2000	76.49	0.0019	1	0.001222458	0.030577269
KDD_US3000	75.33	0.0018	1	0.001165469	0.02963015

ISBN NO: 978-81-908388-7-0

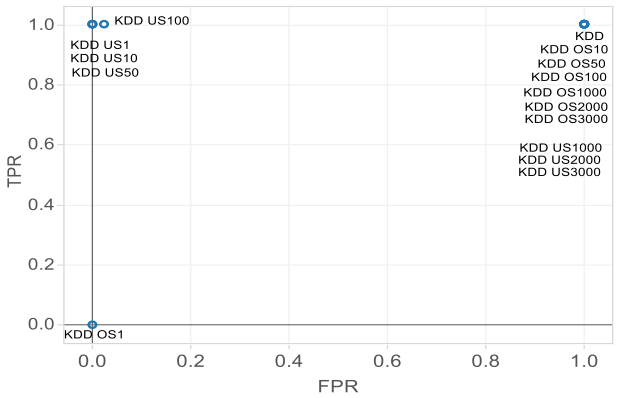

Fig. 6.ROC Plot for KDD CUP 99 oversampled and under sampled data for Random Forest Classifier

PR curve shown in Figure 7 exhibits a state similar to the ROC. The worst scenario is exhibited by oversampled KDD dataset with an imbalance of 1, while high precision and recall rates were exhibited by all the other datasets. Undersampled data (IR: 1-100) exhibits excellent TNR (~1), NPV (0.5-1) and MCC (0.6-1), while other sets exhibit a value of ~0 for all the metrics.

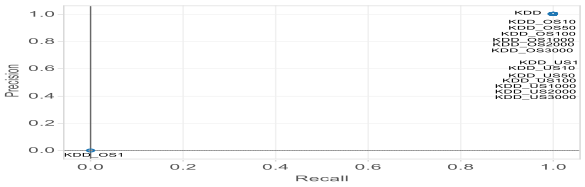

Fig. 7. PR Plot for KDD CUP 99 oversampled and under sampled data for Random Forest Classifier

Figure 7. PR Plot for KDD CUP 99 oversampled and under sampled data for Random Forest ClassifierFigures 8 and 9 presents accuracy and reliability obtained from the KDD dataset exhibiting varied imbalance levels. It could be observed that the imbalance does not affect the accuracy levels (99%) (except for the oversampled data exhibiting imbalance 1). The best reliability rates observed was 66.66% exhibited by the under sampled data with imbalance levels of 1, 10, 50 and 100, while others exhibit a reliability level of 33%.

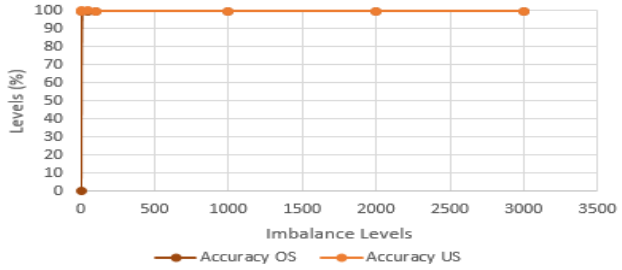

Fig. 8. Accuracy Vs. Imbalance Levels

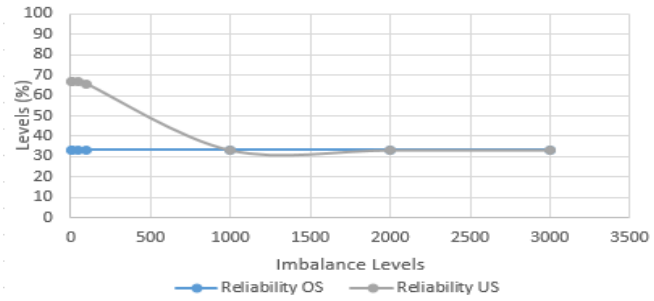

Fig. 9. Reliability Vs. Imbalance Levels

Table 2: Results for Random Forest Classifier

Data Set	Accuracy %	Reliability %	Kappa	True Negative Rate (TNR)	Negative Predictive value	Mathews Correlation coefficient
KDD	99.9	33.33	0	0	0	0
KDD_OS1	0.0288	33.33	0	1	0.000287758	0
KDD_OS10	99.9	33.33	0	0	0	0
KDD_OS50	99.9	33.33	0	0	0	0
KDD_OS100	99.9	33.33	0	0	0	0
KDD_OS1000	99.9	33.33	0	0	0	0
KDD_OS2000	99.9	33.33	0	0	0	0
KDD_OS3000	99.9	33.33	0	0	0	0
KDD_US1	99.9	66.66	-0.0006	1	0.488372093	0.698730868
KDD_US10	100	66.6	-0.0006	1	1	1
KDD_US50	100	66.6	-0.0006	1	1	1
KDD_US100	99.9	65.87	-0.0006	0.97619	1	0.988020134
KDD_US1000	99.9	33.3	0	0	0	0
KDD_US2000	99.9	33.3	0	0	0	0
KDD_US3000	99.9	33.3	0	0	0	0

ISBN NO: 978-81-908388-7-0

The following plots are constructed using a Bank dataset, obtained from a Brazilian Bank. Bank data has an intrinsic imbalance of 25. Naïve Bayes and Random Forest algorithms can withstand until an imbalance level of 100, hence the dataset is used directly without sampling. ROC curve (Figure 10) shows that the Random Forest Classifier exhibits low FPR levels and very high TPR levels, however Naïve Bayes algorithm exhibits a FPR level of 0.5. Both the algorithms exhibit high precision and recall levels, observed from Figure 11. Further, the correct and incorrect levels of classifications in Figure 12 also exhibit that Random Forest classifier performs well on bank data.

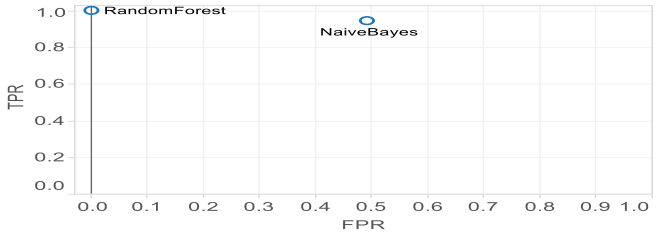

Fig. 10.ROC Plot for Bank Data

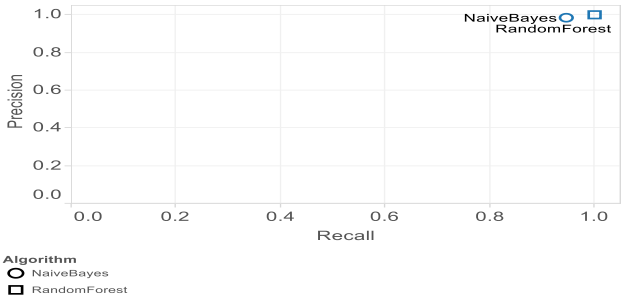

Fig.11.PR Plot for Bank Data

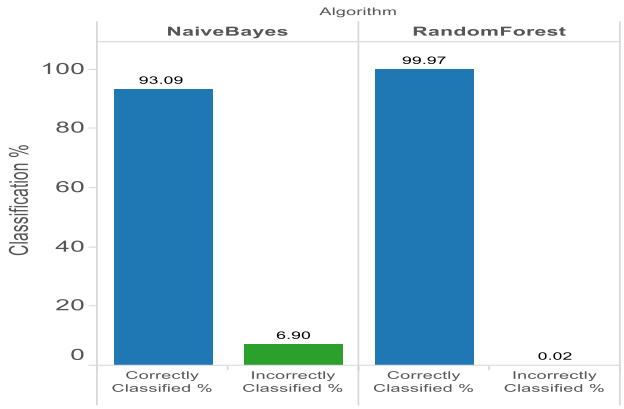

Fig. 12.Classification Levels

Results from Table 3 shows that Naïve Bayes lacks I terms of True Negative Rate (TNR) and Negative Prediction Rate (NPR), while Random Forest exhibits efficient performance in terms of all the metrics.

Table 3: Results for Bank Data

Metric	Naïve Bayes	Random Forest
F-Measure	0.94	1.00
TNR	0.51	1.00
FNR	0.05	0.00
Kappa	0.32	1.00
NPR	0.27	1.00
Accuracy %	93.09	99.07

IV. Discussion

Experiments conducted in this paper are based on KDD for imbalance analysis and Bank Data for analyzing the efficiency of the algorithm in the transaction domain. In order to avoid additional complexities, existing classifiers were used in this paper to identify the impact of imbalance on Big Data. The major concern of the authors is to propose a zero-event anomaly detection/ prevention framework that can identify credit/ debit card frauds as they occur (zero buffer time). The transaction entries pertaining to credit card frauds are imbalanced. Hence an analysis of the effectiveness of algorithms in an imbalanced environment was performed in this paper. Experimental results on KDD dataset reveals that data imbalance has a huge impact on the data even if the number of records to be considered are very high with considerable number of minority entries. Since credit card transactions fall under this category, it becomes mandatory to identify a solution to counter imbalance. It was observed from the experiments that under sampling performs better than oversampling. Hence under sampling component is preferred for the proposed framework.

Experiments were conducted on Bank data without sampling. It was observed that both the techniques performed well, however preference of the authors is towards Random Forest classifier (in terms of handling imbalance) due to its effective prediction of both the classes and low false prediction rates.

V. Conclusion

The advent of Big Data had been to solve the problems existing in real world, but it could be observed that most of the real world problems have huge imbalance levels associated with them. This paper analyses the solutions provided by Big Data with the perspective of Data Imbalance. The current analysis reveals that even after using an effective algorithm, the probability of getting a usable result starts to decrease with increase in the imbalance.

Future works will be based on analyzing some of the most prominent algorithms for their efficiency to work with imbalanced huge data. Analysis will be based on their predictability and their parallelizability. Due to the Big Data environment, parallelizability of an algorithm plays a vital role in improving the performance of algorithms, hence this

property is provided equal importance with respect to the performance of an algorithm. Under sampling techniques will also be explored for usage in the intrusion detection framework. Further, it was also identified that there exist several other factors (apart from imbalance) such as noise, borderline data, effects of anonymization etc., in transaction data that impacts the performance of a classifier. These factors also need to be considered while developing techniques to operate on the data. Future works of the authors will also be based on developing techniques that can operate on these inherent characteristics of transaction data to provide a fast and effective system with improved performance levels for real-time data.

REFERENCES

[1] X. Wu, "Data mining with big data." IEEE transactions on knowledge and data engineering 26.1 (2014): 97-107.

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

[7] 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.

[8] M. Pazzani, C. Merz, P. Murphy, K. Ali, T. Hume, and C. Brunk, "Reducing Misclassification Costs. In Proceedings of the Eleventh International Conference on Machine Learning San Francisco", CA. Morgan Kauffmann, 1994.

[9] 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.

[10] N. Japkowicz, "The Class Imbalance Problem: Significance and Strategies", In Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000): Special Track on Inductive Learning Las Vegas, Nevada, 2000.

[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] M. Kubat, R. Holte, and S. Matwin, "Machine Learning for the Detection of Oil Spills in Satellite Radar Images", Machine Learning, 30, 195–215, 1998.

[13] A. Solberg, and R. Solberg, "A Large-Scale Evaluation of Features for Automatic Detection of Oil Spills in ERS SAR Images", In International Geoscience and Remote Sensing Symposium, pp. 1484–1486 Lincoln, NE, 1996.

[14] 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.

[15] T.M. Ha, and H. Bunke, "Off-line, Handwritten Numeral Recognition by Perturbation Method", Pattern Analysis and Machine Intelligence, 19/5, 535–539, 1997.

[16] L. Breiman. "Random forests." Machine learning 45.1: 5-32, 2001.

[17] M. F. A. Gadi, X. Wang, A.P. do Lago, "Credit card fraud detection with artificial immune system." In International Conference on Artificial Immune Systems (pp. 119-131). Springer Berlin Heidelberg, 2008.

[18] J. Davis, and G. Mark. "The relationship between Precision-Recall and ROC curves." Proceedings of the 23rd international conference on Machine learning. ACM, 2006.

[19] T. Fawcett, "An introduction to ROC analysis." Pattern recognition letters, 27.8 (2006): 861-874.

[20] N. Tomašev, and M. Dunja "Class imbalance and the curse of minority hubs." Knowledge-Based Systems 53 (2013): 157-172.

[21] J. Van Hulse, T. M. Khoshgoftaar, & A. Napolitano, "Experimental perspectives on learning from imbalanced data." In Proceedings of the 24th international conference on Machine learning (pp. 935-942). ACM, 2007.

[22] H. He, and G. Edwardo "Learning from imbalanced data."Knowledge and Data Engineering, IEEE Transactions on 21.9 (2009): 1263-1284.

[23] A. FernáNdez, "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches." Knowledge-based systems 42 (2013): 97-110.

[24] V. López, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics." Information Sciences 250 (2013): 113-141.

[25] M. Galar, "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) 42.4 (2012): 463-484.

[26] G.E. Batista, R.C. Prati, & M.C. Monard, M. C. "A study of the behavior of several methods for balancing machine learning training data." ACM Sigkdd Explorations Newsletter, 6(1), 20-29, 2004.

ISBN NO: 978-81-908388-7-0