ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 3, Issue 9 , September 2016

Fast and Efficient Credit Card Fraud Detection on Real-Time Data Using Spark

A.Sarmila banu, Dr.M.Mohamed Surputheen

Research Scholar, Department of Computer Science, Jamal Mohamed College (Autonomous), Tiruchirappalli, Tamilnadu, India.

Associate Professor, Department of Computer Science, Jamal Moha``med College (Autonomous), Tiruchirappalli, Tamilnadu, India.

ABSTRACT:Fraud Detection is one of the mostly used techniques and is also the one requiring the most updates. An effective fraud detection system is the one that exhibits perfect detection rates with zero false positives. However, the requirement for such a system persists until now. This is due to the increase in the technology, that influences both the ends; the user and the fraudster. Hence it becomes mandatory that the users need to stay a step ahead in this scenario. This paper uses Hadoop architecture to provide a fraud detection technique based on the streaming nature of the data. Naïve Bayes is used as the detection technique and Spark framework is used for implementing the algorithm. Experiments show that the proposed framework exhibits excellent detection levels on the fast streaming data.

KEYWORDS:Credit Card Fraud Detection; Data Imbalance; Anomaly Detection; Naïve Bayes; Spark

I. INTRODUCTION

Dependency on technology has paved way for various advancements in E-Commerce and the M-Commerce domains. Increase in the number of online users has led to an increase in the types of services made available, and this mutual dependency is increasing at a greater rate than ever before. Hence we are facing serious threats in terms of security and privacy. The automation of several essential service sector elements such as insurance, electricity bill payment, telephone bill payments and mobile recharge are due to the emergence of E-Governance projects. This forces people to depend on online and mobile banking. This in turn has resulted in increased number of online fraudulent behaviors such as Credit Card Fraud, Online Identity Thefts, Banking Fraud, Insurance Fraud and Money Laundering.

However, another advancement in technology has made the process of identification and detection of frauds possible. Initially we were not even aware of such fraudulent activities, however the current technologies has the ability to collect, store and process huge amounts of data that are generated by real world processing systems. The rate of data generation was very high and hence the legacy systems were not capable of handling them. With faster and better storage facilities, we are able to collect more data currently than it was collected in the whole of the past. These data can range from Clickstreams, Server Logs, Sensor data, Cameras and other Ubiquitous Computing Devices. The generated data can be structured, semi-structured or even unstructured data, hence the system must be capable of handling this data. The current generation Big Data Platforms allows us to perform all the above mentioned processes. Hence using the current generation technology, we can infer and reason every anomalous pattern which was not even close to a possibility in the past.

However, certain speed bumps also exist in this scenario. Even with such technological advancements, it is impossible to prevent fraud incidents, as it requires techniques that are real time. This paper concentrates on detection and prevention of fraudulent activities through advanced data processing techniques. The huge amount of data that had been the base for performing deep learning and detection of fraudulent transactions, also possess several technological challenges that prevents us from completely utilizing the data for the detection process. The challenges in using huge data ranges from collection to storage to processing.

Further, due to the huge nature of the data, data imbalance also occurs on a multi-dimensional scale. It has become more difficult, as the regular data itself is too huge and the fraudulent data is way too few. Also with big data the imbalance is much higher requiring much better solutions. The unstructured nature of the data makes it further complicated as an appropriate structure is to be imposed upon data prior to processing.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 3, Issue 9 , September 2016

II. RELATED WORKS

Credit card fraud detection technique, being a legacy technique, hence it has several contributions in literature. However, it is also a fast moving technique, hence the older contributions have become dysfunctional as of now. This section discusses some of the most prominent and recent techniques in this area.

Vlasselare et al. presented a credit card fraud detection mechanism that uses several intrinsic features to perform the detection process [1]. It performs fraud detection by identifying intrinsic features identified from the incoming transactions. This technique has its major focus on the buying behaviour of the customers. The major parameters utilized are recency, frequency and monetary levels. These properties are used to predict frauds. Further, several network based features are identified from the transactions and are used to derive a time-dependent suspiciousness score. The suspiciousness score is used to identify the legitimacy of the transaction. Customer spending based fraud detection techniques are currently on the raise. These techniques mostly tend to be unsupervised. Some of the unsupervised fraud detection techniques that work on the basis of customer's spending history are [2-5].

Bolton et al. presented a clustering based technique that analyses spending behaviour of the user for detecting frauds [2]. An alarm is triggered when a transaction violates the regular spending pattern. Weston et al presented a peer group based fraud detection method [3]. This technique uses grouping behavior of the transactions in identifying anomalies. Self-organizing maps is another major grouping technique that can be used for fraud detection process. Ouah et al. [4] and Zaslavsky et al. [5] presents a grouping strategy using SOM to identify fraudulent transactions.

Artificial Neural Networks (ANN) is one of the conventional techniques in the area of fraud detection. This being a machine learning technique, is more suitable for identifying anomalies due to the learning mechanism involved in it. Several contributions for anomaly detection can be found using ANN [6-12, 20].

Hugeness of the involved data has also led to several ensemble based detection techniques. Some ensemble based fraud detection techniques includes random forests [13], SVM [14] genetic algorithms [15] and hidden Markov models [16].

Mahmoudi et al. presented a Modified Fischer Discriminant Analysis based anomaly detection method [17]. This technique uses Fischer Discriminant function to identify anomalies. Further, it also provides more prominence to the minor classes making it more suitable for operating on imbalanced data. Further, this technique also helps in reducing the number of false positives. An Artificial Immune System (AIS) based fraud detection model was presented by Halvaiee et al. in [18]. This technique utilizes AIS to identify legitimate transactions from the fraudulent transactions. Zareapoor et al presented an ensemble based classifier in [19]. This technique utilizes several classifiers, groups their results to identify fraudulent transactions. Though several methods exist for identifying anomalies, most of them do not consider the imbalance nature of data, even if they so, their algorithmic complexities are high.

III.OUR APPROACH

Detecting fraud is always meant to be a continuous process. But due to technological deficits, real time based intrusion detection has not been feasible. Due to the improved storage and processing architectures, this process is now feasible. Hadoop has been the technological breakthrough in bringing about this huge change, by enabling the usage of commodity hardware for Big Data processing. One downside of this approach is that the data uses batch processing, which will not be affordable in real time processing. Several Hadoop related projects have been carried out to enhance the flexibility of the Hadoop system. One such project is Apache Spark, which is a fast and general engine for large scale data processing. The advantage of Spark is that it runs programs that can be 100x faster than the Map Reduce programs in memory and 10x faster in terms of disk. Hence Spark architecture is used for the fraud detection process.

The Naïve Bayes classifier [21] implemented for Classification is parallelized and functions with data in-memory, rather than from the hard disk (Figure 1). Due to the presence of unstructured data, the parallelized Naïve Bayes algorithm is implemented in a generic manner. Since the data format can vary, the first phase performs input analysis and cleaning. The input corresponding to both training and testing phases are initially processed. This processing is also

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 3, Issue 9 , September 2016

parallelized to make sure that it does not create a bottleneck for the subsequent processes. The input data is analyzed and is divided into the format accepted by the Classifier algorithm. Besides, the real time data, generated is prone to inaccuracies and missing values. These values are adjusted and converted to corresponding formats in this phase.

Data for Classification is in general divided into test and training data. This stage performs the process of dividing the data into test and training sets. Parallelized random splitting of data is used, which makes sure that certain classes are not concentrated in specific regions. Data is first randomized and then the division of data is carried out. Most classifier applications use a division ration of 7:3 for the training and test datasets respectively, hence out default segregation mechanism uses these values for dividing the dataset. It is also possible to change this ratio according to the user's convenience. This phase is not mandatory, since several data sets contain their own set of training and test data. In such a scenario, this stage can be skipped and the classifier training can be carried out.

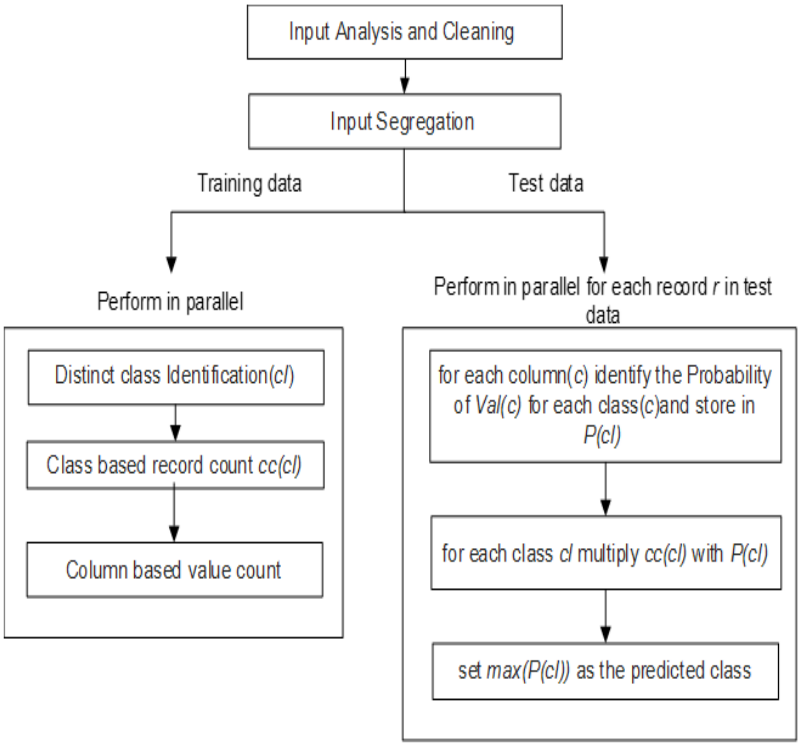

Fig. 1. Parallel Naïve Bayes Classifier for credit card Fraud Detection

The process of classification is carried out on the training data. This paper uses a parallel Naïve Bayes algorithm that works in-memory to provide results in real time. Naïve Bayes is a conditional probability model that classifies a certain instance represented by a vector X=(x1, x2,….xn) with n features to a define class from the available set of classes C=(c1, c2….cm). The class assignment is carried out by identifying the probability for every class p(ck|x1,x2….xn). The conditional probability is represented as,

$$p(C_k|X) = \frac{p(C_k)p(X|C_k)}{p(X)} \quad (1)$$

Though this method is the simplest form of Classification, it could be observed that this method is free from additional dependencies such as requirements for data normalization and can incorporate any type of data (integers, real numbers or strings). Hence it becomes the most valid candidate for usage in Big Data scenario, where several data structures can be expected. Another major advantage of this method is that from the experiments, it was observed that

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 3, Issue 9 , September 2016

the method works particularly well on imbalanced data. Data imbalance has always remained a pressing issue when it comes to classification. It is always said to have a negative impact on the classified results.

IV. EXPERIMENTAL RESULTS

Experiments were carried out in Cloudera VM using Hadoop 2 and Spark 1.4. Brazilian Bank dataset containing ~345 K records is used for analysis. The dataset has an imbalance ratio of 25, containing anonymized transaction details of the customers.

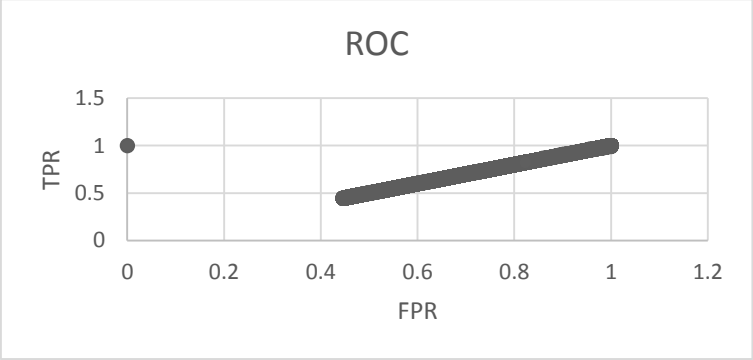

Fig. 2. ROC

The ROC plot for bank dataset operated on Naïve Bayes classifier is presented in Figure 2. It could be observed that the TPR values range from moderate to high, representing good prediction levels of the algorithm. However, the false positive rates were also found to be high. Hence it could be observed that the current technique also exhibits a high false positive rate, which needs to be addressed in the further contributions.

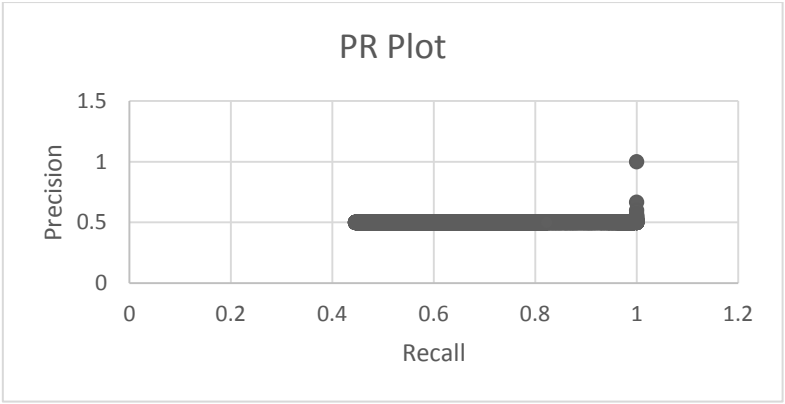

Fig. 3 PR

The PR plot representing precision and recall values are presented in figure 3. It could be observed that the classifier exhibits moderate precision levels and moderate to high recall levels. This exhibits efficient retrieval rates of the Naïve Bayes classifier.

V. CONCLUSION

Occurrence of fraud in transactions is one of the major concerns of any system performing monetary transactions. Hence anomaly detection in credit/ debit card transactions has become inevitable for any system handling online financial transactions. This paper presents a Spark based streaming fraud detection technique for credit card fraud

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 3, Issue 9 , September 2016

detection. However, several shortcomings were observed, such as high false positive rates. Future contributions will be based on hybridizing or modifying the algorithms to eliminate false positives and provide higher precision levels.

REFERENCES

[1] V. Van Vlasselaer, C. Bravo, O. Caelen, T. Eliassi-Rad, L. Akoglu, M. Snoeck, and B. Baesens, "APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions." Decision Support Systems, 75, pp.38-48,2015.

[2] R.J.Bolton, D.J. Hand, "Unsupervised profiling methods for fraud detection." Proceedings of the VII Conference on Credit Scoring and Credit Control, pp. 235–255,2001. (Edinburgh, United Kingdom).

[3] D.J.Weston, D.J. Hand, N.M. Adams, C. Whitrow, P.Juszczak, "Plastic card fraud detection using peer group analysis." ADAC 2 (1) 45–62, 2008.

[4] J.T. Quah, M. Sriganesh, "Real-time credit card fraud detection using computational intelligence." Expert Syst. Appl. 35 (4) 1721–1732, 2008.

[5] V.Zaslavsky, A. Strizhak, "Credit card fraud detection using self-organizing maps." Inf. Secur. 18- 48, 2006.

[6] E.Aleskerov, B. Freisleben, B. Rao, "Cardwatch: a neural network based database mining system for credit card fraud detection" Computational Intelligence for Financial Engineering (CIFEr), Proceedings of the , pp. 220–226,1997.

[7] R.Brause, T.M. Langsdorf, M. Hepp, "Neural data mining for credit card fraud detection." Proceedings. 11th IEEE International Conference on Tools with Artificial Intelligence, pp. 103–106,1999.

[8] J.R.Dorronsoro, F. Ginel, C. Sgnchez, C. Cruz, "Neural fraud detection in credit card operations." IEEE Trans. Neural Netw. 8 (4) 827–834,1997.

[9] S.Ghosh, D.L.Reilly, "Credit card fraud detection with a neural-network." Proceedings of the Twenty-seventh International Conference on System Sciences, vol. 3, pp. 621–630,1994.

[10] S.Maes, K.Tuyls, B. Vanschoenwinkel, B. Manderick, "Credit card fraud detection using bayesian and neural networks." Proceedings of the 1st International NAISO Congress on Neuro Fuzzy Technologies,2002.

[11] A.Shen, R.Tong, Y. Deng, "Application of classification models on credit card fraud detection." Service Systems and Service Management, International Conference on. pp. 1–4,2007.

[12] M.Syeda, Y.Q. Zhang, Y. Pan, "Parallel granular neural networks for fast credit card fraud detection" Proceedings of the IEEE International Conference on Fuzzy Systems, vol. 1, pp. 572–577,2002.

[13] K. Henderson, B. Gallagher, L. Li, L. Akoglu, T. Eliassi-Rad, H.Tong, C. Faloutsos, "It's who you now: graph mining using recursive structural features." Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM , pp. 663–671,2011.

[14] S. Bhattacharyya, S. Jha, K. Tharakunnel, J.C.Westland, "Data mining for credit card fraud: a comparative study." Decis. Support. Syst. 50 (3) 602–613,2011.

[15] E. Duman, I. Elikucuk, "Solving credit card fraud detection problem by the new metaheuristics migrating birds optimization." in: I. Rojas, G. Joya, J. Cabestany (Eds.), Advances in Computational Intelligence. Vol. 7903 of Lecture Notes in Computer Science, Springer, Berlin Heidelberg, pp. 62–71,2013.

[16] A. Srivastava, A. Kundu, S. Sural, A.K.Majumdar, "Credit card fraud detection using hiddenmarkov model." IEEE Trans. Dependable Secure Comput. 5 (1) 37–48,2008.

[17] N. Mahmoudi, E.Duman, "Detecting credit card fraud by Modified Fisher Discriminant Analysis."Expert Systems with Applications, Volume 42, Issue 5, Pages 2510-2516,2015.

[18] N.S. Halvaiee, and M.K.Akbari, "A novel model for credit card fraud detection using Artificial Immune Systems." Applied Soft Computing, 24, pp.40-49,2014.

[19] M. Zareapoor, P. Shamsolmoali, "Application of credit card fraud detection: Based on bagging ensemble classifier."Procedia Computer Science, Volume 48, Pages 679-685,2015.

[20] J. West, M. Bhattacharya. "Payment Card Fraud Detection Using Neural Network Committee and Clustering." Computers & Security, Volume 57, Pages 47-66, 2016.

[21] P.Bradley, and T.Louis, "Bayes and Empirical Bayes Methods for Data Analysis". London: Chapman & Hall,1996.