

Predictive Analysis of Transaction Fraud leveraging Big Data for Deep Learning

Priyanka Purushu, Junghoon Heo, Sooyoung Kim, Yeon Pyo Kim and Jongwook Woo*
AT&T, US
Softzen Co. Ltd., Republic of Korea
Department of Information Systems, California State University Los Angeles, US
[e-mail: priyanka.purushu89@gmail.com, {jhheo, sykim, ypkim}@softzen.co.kr,
jwoo5@exchange.calstatela.edu]
*Corresponding author: Jongwook Woo

Abstract

The paper presents distributed deep learning using Spark Big Data platform for Financial Fraud Detection on transactional activity. It is scalable deep learning systems, where Deep Learning is integrated into the Spark Big Data platform. That is, the Big Data platform can be used for data cleaning and analysis to store and compute massive data set, and then, the input data set is loaded to the model without the latency for training with the computing tasks of the distributed deep learning. With this scalable deep learning, deep learning models is built for fraud detection and Spark machine learning classifiers in the paper. Two deep learning algorithms are used as Feed Forward neural network: one from Intel and another from Apache Spark, as the traditional Random Forest, Decision Tree, and Logistic Regression, which run in Big Data platform. It is shown that with the same data set, scalable deep learning on Spark presents the prediction model with the best accuracy in Recall under the similar computing time. In summary, the paper shows that integrating Deep Learning into the Big Data platform for fraud detection is acceptable, and it presents the best performance as well, especially in Recall.

Keywords: Fraud Classification, Big Data, Deep Learning, Predictive Analysis, Distributed Computing, Feed Forward, Scalable Deep Learning

1. Introduction

Woo et al. defines Big Data as a non-expensive supercomputer composed of commodity servers. It is a distributed parallel computing systems to compute and store a large-scale data. A large-scale data means data of giga-bytes or more, which cannot be processed well or too expensive using traditional computing systems [3, 13].

The traditional MapReduce in Big Data Hadoop solution has a scheduling overhead and it lacks iterative computation because of the intermediate data at the storage, which slows down its performance on machine learning. However, Spark supports in-memory computing, so it is efficient at iterative computations and thus popular for the development of large-scale machine learning applications [10].

Financial Fraud can be a devastating issue with extensive ramifications on any business, finance industry, corporate and government segments, and for individual consumers [4]. With technological advancements, these transaction frauds are becoming more intricate. Today in the data-driven world, the fraudulent transactions can be tracked down by analyzing the massive transaction data set with the use of Big Data platforms and data mining approaches.

While carrying research on this topic, challenges are encountered in finding a dataset on financial

This research was supported as a Brain Pool Program of 2020 by The Ministry of Science and ICT and National Research Foundation of Korea and a research credits program from Google Cloud Platform.

fraud detection. These kinds of financial datasets are not publicly available due to the nature of the information. Our work can be related on the same lines, just like other researchers. For us, finding the dataset was not much difficult due to the availability of PaySim's synthetic dataset. A synthetic transactional data was developed by the PaySim simulator which incorporated both: normal customer behavior and fraudulent behavior [9].

The paper aims at doing predictive analysis on the target value, which is column "isFraud" and detect if a money transaction is a fraud or not. The dataset size is approximately 470MB and it has eleven features.

It is acknowledged that the 470MB is not Giga- or Tera-bytes of the massive data set. However, it would be better to adopt the Spark Big Data architecture and develop predictive models, which is linearly scalable to compute massive data set by adding more spark nodes to the cluster concerning the data set. In addition to this, Spark-in-memory processing in Python is much faster than the traditional sequential Python approach.

The paper is to predict if a financial transaction is a fraud or not using classification models. In this paper, the data is analyzed by integrating two machine learning platforms: Apache Spark ML and Deep Learning (DL). Apache Spark ML is Big Data platform as distributed parallel computing systems. It is helpful to leverage the Big Data platform adding Deep Learning libraries from Apache Spark and Intel, which allows scalable deep leaning.

The paper is composed of the sections: Related Work, Financial Transaction Data, Big Data Predictive Analysis with Machine / Deep Learning, Experimental Result, and Conclusion.

2. Related Work

The research works related to financial fraud transactions is the popular area with the implementation of machine learning models.

Kamaruddhin et al. built a classification model using Auto-Associative Neural Network to detect credit card fraud. It runs in Spark platform with a hybrid architecture of Particle Swarm Optimization [8]. They were able to achieve an accuracy of 89%. However, unlike us, they worked on a comparatively smaller dataset of 291.7MB in size that contains only 9 features.

Hormozi et al. developed a financial Fraud model on the cloud platform to detect a credit card fraud using an Artificial Immune System's algorithm [6]. The model is based on the negative selection algorithm. They adopted MapReduce computing engine of Apache Hadoop on a dataset with 300,000 rows. It is comparatively smaller than our dataset, which has 6,362,620 rows.

Pryanka used Spark MLlib and Analytics Zoo provides fast, distributed implementations of deep learning model, feedforward, as well as standard models of the legacy data science algorithms, including Random Forest, Decision Tree, and Logistic Regression in Spark cluster. They compare the accuracy and computing time with the traditional sequential machine [1].

In this paper, Pryanka's approach [1] is extended by adding deep learning classification models: Feed Forward in BigDL and Multilayer Perceptron Classifier in Spark MLib, and compare the accuracy and computing time with the legacy Big Data models in Spark cluster.

3. Financial Transaction Data

It is not easy to collect or find the data set for the transaction fraud detection. So we synthesized dataset using the simulator PaySim [9]. PaySim can generate a synthetic dataset from the existing private dataset. Once the data set is generated, we can build a model with the data to evaluate the performance of fraud detection methods. That is, We can simulate the regular operation of transactions and injects malicious behavior with the data set. For this paper, we collect data set from PaySim, which is based on the data at a mobile money service in an African country and simulates mobile money transactions from one month of financial logs.

The data has a size of 470 MB with 6,362,620 rows. The dataset contains 10 attributes and the target column is 'isFraud'. A transaction can either be non-fraudulent, indicated by a 0, or

fraudulent, marked by a 1, which makes this to a binary classification problem.

A sample row of the dataset looks like: (1, PAYMENT, 1060.31, C429214117, 1089.0, 28.69, M15916 54462, 0.0, 0.0). And the attributes of the dataset with metadata has been explained in further detail below:

• Transaction Type: CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER.
• Amount: the amount of the transaction in local currency.
• nameOrig: the origin of the transaction.
• oldbalanceOrg: initial balance at the origin before the transaction.
• newbalanceOrig: new balance at the origin after the transaction.
• nameDest: the destination of the transaction.
• oldbalanceDest: initial balance of the destination before the transaction.
• newbalanceDest: new balance of the destination after the transaction.
• Step: The unit of time. The total steps are 744 for one month simulation.
• isFraud: The label to show if the transaction is fraud or not. The flag attribute is either 0 or 1.

The dataset provides 5 numeric attributes (amount, oldbalanceOrg, newbalanceOrg, oldbalanceDest, newbalanceDest), 3 categorical attributes (step, type, isFraud) and two string attributes (nameOrig, nameDest).

4. Big Data Predictive Analysis with Machine / Deep Learning

Apache Spark supports in-memory processing as a distributed parallel computing systems. It supports machine learning libraries. We uses Hadoop Big Data systems, which is linearly scalable. It is composed of HDFS file systems, Spark computing engine , and YARN resource management.

For classifying and detecting the Fraud in the financial data set, three traditional machine learning algorithms are considered: Random Forests, Decision Tree, and Logistic Regression. Logistic Regression is to find out and estimate the specific number of values and a popular method to predict a categorical response with the probability of the outcomes. Decision Tree is an analytical tool that supports decision making by including event outcomes or their possible consequences.

Classifications are commonly used to detect ad click and credit card Fraud. Decision Trees is one of the algorithm for classification [11]. It does not require feature scaling and easy to interpret and capture non-linearities and feature interactions. Random Forest is ensembles of decision trees by combining many decision trees with being expressed as a set of de-correlated decision trees. Thus, it reduces the risk of overfitting. The example of Random Forest can be a data set that contains different random values and their class. Then the data set is divided into a lot of subsets with random values and random classes. After the division, the algorithm decides and allocates different classes to each of the independent forests.

For deep learning, feed-forward neural network is adopted, which produces many popular Convolution Neural Networks (CNN). It composes the neural network by copying the connectivity patterns of the neurons from the animal's visual cortex. The deep learning algorithms are built in Spark platform, which are implemented by Apache Spark and Intel's BigDL.

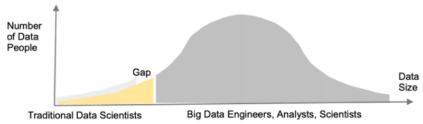

Fig. 1. A gap between the traditional Data Scientists / Deep Learning professionals and the Big Data professionals

The traditional data science develops machine learning models in Python and R for the small dataset. It has the data size of up to Mega-Bytes and generates memory issues when to process Giga-Bytes of data set. Figure 1 shows the gap between the traditional and Big Data. Even the deep learning system with a single server has the similar issue when to read large scale data. Data organization needs more Big Data Engineers, Analysts, and Scientists while the data grows exponentially.

The traditional approach has a limitation to implement deep learning and machine learning models when there is large scale data higher than Giga-bytes. Spark Hadoop cluster is scalable systems by adding more servers when data grows. It means it is linearly scalable. We can leverage the Big Data platforms to read data stored in HDFS. Then, Spark can process iterative computation to build the machine learning models.

5. Experimental Result

The models are developed using the Spark Machine Learning library (Python 2.7.14, Spark 2.3.4) of Dataproc cluster in **Google Cloud Platform**. The cluster is composed of 6 nodes using **n1-standard-64** (64 vCPUs, 240 GB memory, 257 TB storage): 1 Master and 5 Workers. A vCPU is implemented as a single hardware Hyper-thread on either Intel Xeon or ARM EPYC Rome CPU platforms.

Five classification models are implemented based on the three traditional machine learning algorithms - Decision Tree (DT), Random Forest (RF), Logistic Regression (LR). And, the other two models are implemented based on the deep learning algorithm Feed Forward (FF) such as Multilayer Perceptron FF (MFF) and BigDL FF (BFF).

The first experiment with these models takes the input data set "paysimFraudSmoted.csv", which has 6.4 million records, and it has only 6.7% frauds. To improve the performance of accuracy, the data of Fraud and non-frauds is balanced adopting undersampling. It randomly samples non-fraud records that is a majority class. It reduces the entire training data set, which can improve performance to build models. Therefore, 0.4 million non-fraud records are sampled to balance with the raw fraud records: [Fraud : Not-Fraud] = [418,863 : 424,146]

Table 1 shows the experimental result of the experiment using the five classification models in the Spark cluster. MFF is generalized with Cross-Validation (CV) and Train Split Validation (TSV) and presented as MFF CV and MFF TSV, respectively.

Feed-Forward models show exciting results. MFF has almost perfect Recall: 1, where FN is too small about 10 and 71 cases for TSV and CV, respectively, out of 244 K test data. Thus, it can apply to predict TP. BFF has a similar number of TP, FP, TN, FN, so that Recall and Precision are 59% and 52%, but AuROC is 1. Even though AuROC is 1, it may not be applicable.

Table 1. Comparison of Classification Models without balance

Model	Precision	Recall	AUC	Time (mins)
DT	0.976	0.975	0.976	3
RF	0.946	0.860	0.979	13
LR	0.946	0.860	0.905	3
MFF TSV	0.694	1	0.782	2
MFF CV	0.695	1	0.783	4
BFF	0.593	0.516	1	4

Besides FF models, RF has a high AUC (Area under ROC): 0.979 as shown in Table1. Table 2 is the confusion matrix of DT, where DT has the highest Recall, 0.975, and Precision 0.976.

Table 2. Confusion Matrix of DT

	Actual Positive	Actual Negative
Predicted Positive	122,190	3,040
Predicted Negative	3,069	123,954

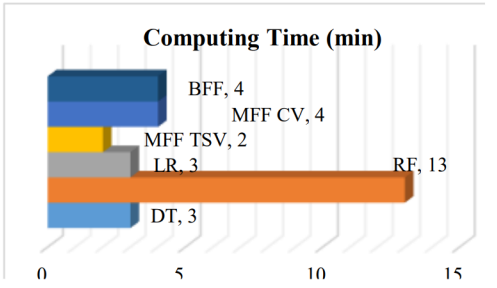

Fig. 2. Performance Comparison of classification models with balanced data

Fig. 2 shows the computing time as performance. MFF TSV has the fast computing time of about 2 minutes while LR, DT, MFF CV, and BFF have similar computing times: 3 - 4 minutes, and RF has a much longer computing time: 13 minutes.

Recall is an important measurement for the accuracy of our fraud detection. Besides, the fast computing time to build a model is another important factor. Therefore, DT or MFF should be preferable algorithms.

6. Conclusion

The Big Data distributed parallel computing systems are shown by integrating Deep Learning into Spark Machine Learning systems. It is to prove that deep learning can run in Big Data as a part of the Spark platform and as the third party solution. The third-party solution is BigDL, developed by Intel. A dataset containing fraudulent and non-fraudulent financial transactions is investigated, which made it to a binary classification problem to predict frauds. Since the dataset was about 470 MB, big data service using Google GCP is adopted to compute the entire data set in a much faster way. At the same time, it reduces computing time in several hours to predict values than traditional systems.

The raw data set is sampled to balance it with a similar ratio, 50% vs. 50%, between Fraud and non-fraud data set because it has better accuracy than a non-balanced raw data set. For the balance data set, the Decision Tree Classifier scored the best precision accuracy with 97.6% and recall with 97.5%. The Feed Forward DL Classifier, Multilayer Perceptron, achieved the best recall accuracy with 99.9%. They compute the accuracy in about 3 - 4 minutes. The Random Forrest classifier presents 97.9% in Area Under ROC, which calculates the accuracy in 13 minutes.

From this, it is concluded that FF Multilayer Perceptron model is the best recall accuracy in predicting the least incorrect non-fraudulent transaction when they are actual non-fraudulent. The next is DF model with the best Precision. RF model is the best accuracy of Area Under ROC, but its Recall is lower than DT and FF Multilayer Perceptron. Recall as the target accuracy measurement is taken because it is the goal to find out which transaction is Fraud. Thus, either DT or FF Multilayer Perceptron algorithms should be accepted for financial transaction fraud classification.

References

[1] Purushu P., and Woo J. 2020. Financial Fraud Detection adopting Distributed Deep Learning in Big Data. KSII The 15th Asia Pacific International Conference on Information Science and Technology (APIC-IST) 2020, July 5 -7 2020, Seoul, Korea, pp271-273, ISSN 2093-0542

[2] Purushu, P., Melcher, N., Bhagwat B., and Woo, J. 2018. Predictive Analysis of Financial Fraud Detection using Azure and Spark ML. Asia Pacific Journal of Information Systems (APJIS), VOL.28, NO.4, pp308~319

[3] Woo, J., and Xu, Y. 2011. Market Basket Analysis Algorithm with Map/Reduce of Cloud Computing, The 2011 international Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2011), Las Vegas.

[4] Financial Transactions & Fraud Schemes (n.d). Retrieved 2018 from https://www.acfe.com/financial-transactions-and-fraud-schemes.aspx

[5] Han, J., and Kamber, M. 2006. Data Mining: Concepts and Techniques, Second edition, Morgan Kaufmann Publishers, 2006, pp. 285–464

[6] Hormozi, H., Akbari, M. K. , Hormozi, E. , and Javan, M. S., 2013. Credit cards fraud detection by negative selection algorithm on hadoop (To reduce the training time), The 5th Conference on Information and Knowledge Technology, pp. 40–43, 2013

[7] Jones, T. A. 2002. Writing a good paper. IEEE Trans. On General Writing, Vol. 1, no. 2, pp.1-10

[8] Kamaruddhin, S. K., and Ravi, V. 2016. Credit Card Fraud Detection using Big Data Analytics: Use of PSOAANN based One-Class Classification. ICIA-16 Proceedings of the International Conference on Informatics and Analytics 2016. Article No. 33

[9] Lopez-Rojas, E. A., Elmir, A., and Axelsson, S. 2016. PaySim: A financial mobile money simulator for fraud detection. The 28th European Modeling and Simulation Symposium-EMSS

[10] Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D.B., Amde, M., Owen, S., and Xin, D.

2015. MLlib: Machine learning in apache spark. arXiv preprint arXiv:1505.06807

[11] Anuj, S., and Prabin, P. K. 2013. A Review of Financial Accounting Fraud Detection based on Data Mining Techniques. International Journal of Computer Applications February 2012

[12] Synthetic Financial Datasets for Fraud Detection (n.d). Retrieved from https://www.kaggle.com/ntnu-testimon/paysim1

[13] Woo, J. 2013. Market Basket Analysis Algorithms with MapReduce. DMKD-00150, Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery, Volume 3, Issue 6, pp445-452, ISSN 1942-4795.

[14] Gupta N, Le H., Boldina M.., Woo, J. 2019 Predicting Fraud of AD click using Traditional and Spark ML. KSII The 14th Asia Pacific International Confer-ence on Information Science and Technology (APIC-IST), pp24-28

[15] García-Gil D, Ramírez-Gallego S., and García S.. 2017. A comparison on scalability for batch big data processing on Apache Spark and Apache Flink. Big Data Anal 2, 1 doi:10.1186/s41044-016-0020-2

[16] Nattar Kannan, S. Sivasubramanian, M. Kaliappan, S. Vimal & A. Suresh 2019. Predictive big data analytic on demonetization data using support vector machine" Cluster Computing. https://doi.org/10.1007/s10586-018-2384-8