2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Anomaly Detection on Big Data in Financial Markets

Mohiuddin Ahmed, *Member*, *IEEE*
Department of ICT and Library Studies
Canberra Institute of Technology, Australia
m.ahmed.au@ieee.org

Nazim Choudhury, Shahadat Uddin
Complex Systems Research Group
The University of Sydney, Australia
nazim.choudhury@sydney.edu.au
shahadat.uddin@sydney.edu.au

***Abstract*—In the modern financial market, market participants use big data analytics to gain valuable insight on historical market data for better decision making. Complying with the three vs (i.e., velocity, volume and variety) of big data, the financial market is considered as a complex system comprised of many interacting high-frequency traders those make decisions based on the relative strengths of these interactions. Researchers have put substantial scholarly input to deal with these anomalies. From the big data perspective, anomaly detection in financial data has widely been ignored despite many organisations store, process and disseminate financial market data for interested customers to assist them to make informed decision abd create competitive advantages. Considering the presence of anomalies in voluminous data from myriad data sources may generate catastrophic decision through misunderstandings of market behaviour. Therefore, in this study, we applied a standard set of anomaly detection techniques, used in big data based on nearest-neighbours, clustering and statistical approaches, to detect rare anomalies present within the historical daily trading information for five years (i.e., 2009-2013) for each stock listed on the Australian Security Exchange (ASX). We also measured the performance of these anomaly detection techniques using a number of metrics to highlight the best performing algorithm. The experimental results suggest that the LOF(Local Outlier Factor) and CMGOS(Clustering-based Multivariate Gaussian Outlier Score) are the best performing anomaly detection techniques.**

***Index Terms*—Anomaly Detection, Financial Markets, Financial Big Data.**

I. Introduction

Data proliferation and increasing technological complexities have transformed the modern financial market where market participants use big data analytics to gain valuable insight of historical market data for better decision making. Voluminous content and unstructured information collected from the stock message boards assist most organizations in gaining new insights of financial data and making decisions by utilizing historical time series information of stocks and commodities. Financial markets generate massive amount of data that follows the three properties (i.e., volume, velocity and variety) of big data. For example, the New York Stock Exchange (NYSE) generates one terabyte of information during each day; high-frequency stock trading algorithms reflect market changes within microseconds; and market data provides price and trade related various information that allow traders or investors to know latest price, amount and volume of transactions and observe historical trends.

The potential value of financial big data is unlocked if it is leveraged to drive decision making. Therefore, traders and organizations need efficient and flawless processing of high volume, rapidly changing diverse data to generate meaningful insights and evidence-based decision making [1]. The core of large financial datasets reflects the extremely detailed combination of decisions generated by the market actors [2]. In some combinations, these myriad decisions, emerging from heterogeneous market actors, can lead to catastrophic conclusions due to the presence of anomalies in the recorded data or the malfunctioning of the systems recording data for the traders and scholars. Since the financial crises in 2007, researchers, risk, and financial practitioners are keener to combine practical and theoretical knowledge of financial markets since in this domain, in one hand, there exists competing theories of commodity pricing, and on the other hand, empirical observations sometimes undermine the confidence of the theoretical belief. Researchers have been attempting to detect stock market anomalies for quite a long time. Yavrumyan pointed out three types of market anomalies; namely (i) technical, (ii) fundamental and finally (iii) calendar [3]. A comprehensive list of market anomalies is described in [4]. In this study, instead of considering market anomalies, as described above, we consider big data perspective. The underlying reason behind this is that big data has become a business imperative and provides solutions to challenges for financial markets. According to a study performed by IBM in 2013 [5], 71% of banking and financial markets are leveraging big data analytics to create competitive advantages for their organizations compared to 36% in 2010, representing a huge 97% increase in just two year. Although substantial amount of research efforts have considered the market anomalies from financial markets theoretical perspective [6], however; it is a research requirement to analyse the stock market anomalies in regards to big data perspective. Recently, a number

ASONAM '17, July 31 - August 03, 2017, Sydney, Australia

ACM ISBN 978-1-4503-4993-2/17/07.../$15.00
http://dx.doi.org/10.1145/3110025.3119402

studies have proposed big data anomaly detection techniques in various domains like distribution grids, cellular networks, however; financial markets are widely ignored. Therefore, in this study, we examine historical daily trading information for the period 2009-2013 (inclusive) for each stock listed on the Australian Security Exchange (ASX), maintained by Securities Industry Research Centre of Asia-Pacific (SIRCA), from big data perspective and apply different unsupervised anomaly detection techniques to detect the presence of rare anomalies within the information provided. Rest of the paper is organized as follows. Section II provides a detailed discussion on the anomaly detection techniques used in the paper. Section III provides the experimental analysis including the dataset used in this study and key insights obtained. Section IV concludes the paper.

II. Preliminary Discussion on Anomaly Detection

Anomaly (also known as outlier) detection is an important data analysis task that detects anomalous or abnormal data from a given dataset. Anomaly detection has been widely studied in statistics, machine learning and financial domain [7], [8]. Three dominant types of anomaly detection approaches include: Supervised, Semi-Supervised and Unsupervised. Supervised and Semi-Supervised approaches both depend on the knowledge from labelled data. The majority of the anomaly detection techniques are supervised in nature which relies on labelled training data. This training data is typically expensive to produce. Moreover, these methods have difficulty in detecting new types of anomalous points in the dataset. Semi-supervised techniques take advantage of a small amount of labelled data and create a model for detecting anomalies. On the other hand, unsupervised anomaly detection techniques do not require any training data and capable of detecting previously unseen anomalies. In recent years, the traditional philosophy of using a knowledge base or external supervision is superseded by unsupervised anomaly detection. Given below are the anomaly detection techniques, described by Ahmed et al. in their paper [9] and used in this study for the purpose of detecting rare anomies in financial data:

- *k-NN*: Each data instance is given a score for being anomalous based on the average distance to the nearest-neighbours.
- *LOF*: LOF provides anomaly score to the data instances based on the local density of the data points.
- *COF*: The connectivity based outlier factor is a modification of the *LOF* approach which can handle outliers deviating from low density patterns.
- *aLOCI*: Calculates the outlier score based on local correlation integral.
- *LoOP*: The LoOP score represents the probability that an instance is a local density outlier.
- *INFLO*: Calculates the outlier score based on Influenced Outlierness, proposed by Jin et al [10].
- *CBLOF*: CBLOF creates clusters from the given dataset and categorizes the clusters into small clusters and large clusters. The anomaly score is then calculated based on the size of the cluster an instance belongs to as well as the distance to the nearest large cluster centroid.
- *LDCOF*: This local density based anomaly detection algorithm sets the anomaly score based on the distance to the nearest large cluster divided by the average cluster distance of the large cluster.
- *CMGOS*: This method calculates the anomaly score based on a clustering result. The outlier score of an instance is dependent on the its distance to the cluster center.
- *RPCA*: Principal component analysis is another commonly used technique for detecting subspaces in datasets that may serve as an anomaly detection technique, such that deviations from the normal subspaces may indicate anomalous instances.
- *LIBSVM*: Computes the outlier score using one-class SVMs (Support Vector Machines). This operator extends the semi-supervised one-class SVM such that it can be used for unsupervised anomaly detection.

III. Experimental Analysis

A. Evaluation Measures for Anomaly Detection

Generally in data mining and machine learning research domain and specifically for anomaly detection evaluation, *Precision* and *Recall* are the two important and widely used metrics [7]. Another important metric involved in evaluation of anomaly detection is *False Positive Rates* or *FPR*. *FPR* indicates the percentage of instances that were inaccurately detected as anomalous among all instances that are identified as anomalous. Also, there is a measure that combines *Precision* and *Recall*, as the harmonic mean of these two metrics, called *F-measure*.

B. Dataset Analysis

In this study, we extracted five years (2009-2013 inclusive) of historical daily trading information for each stock listed on the Australian Stock Exchange (ASX), provided by SIRCA. Each entry in the dataset provide daily transaction information for a stock including the date of the transaction, identification of the stock (i.e., stock code), highest and lowest trade value for the date, last trading price of the stock for the day, volume (i.e., quantity) and total monetary value of the transactions for the day. In Table 1, we provide the basic statistics of the dataset.

There exists certain challenges in anomaly detection process. For example, defining a normal region that encompasses every possible normal practice and simultaneously, the precise boundary between normal and anomalous behaviour is very difficult. Therefore, it is challenging to define the exact notion of anomaly which is domain variant. In case of stock exchange data in the financial domain, presence of zero values in regards to transaction is erroneous since it conveys misleading information about a true transaction. Since the dataset provides us daily historical information considering five variables, (i.e., daily high, daily low, last trading value for the day, total quantity of stock transactions and total price worth of the stock

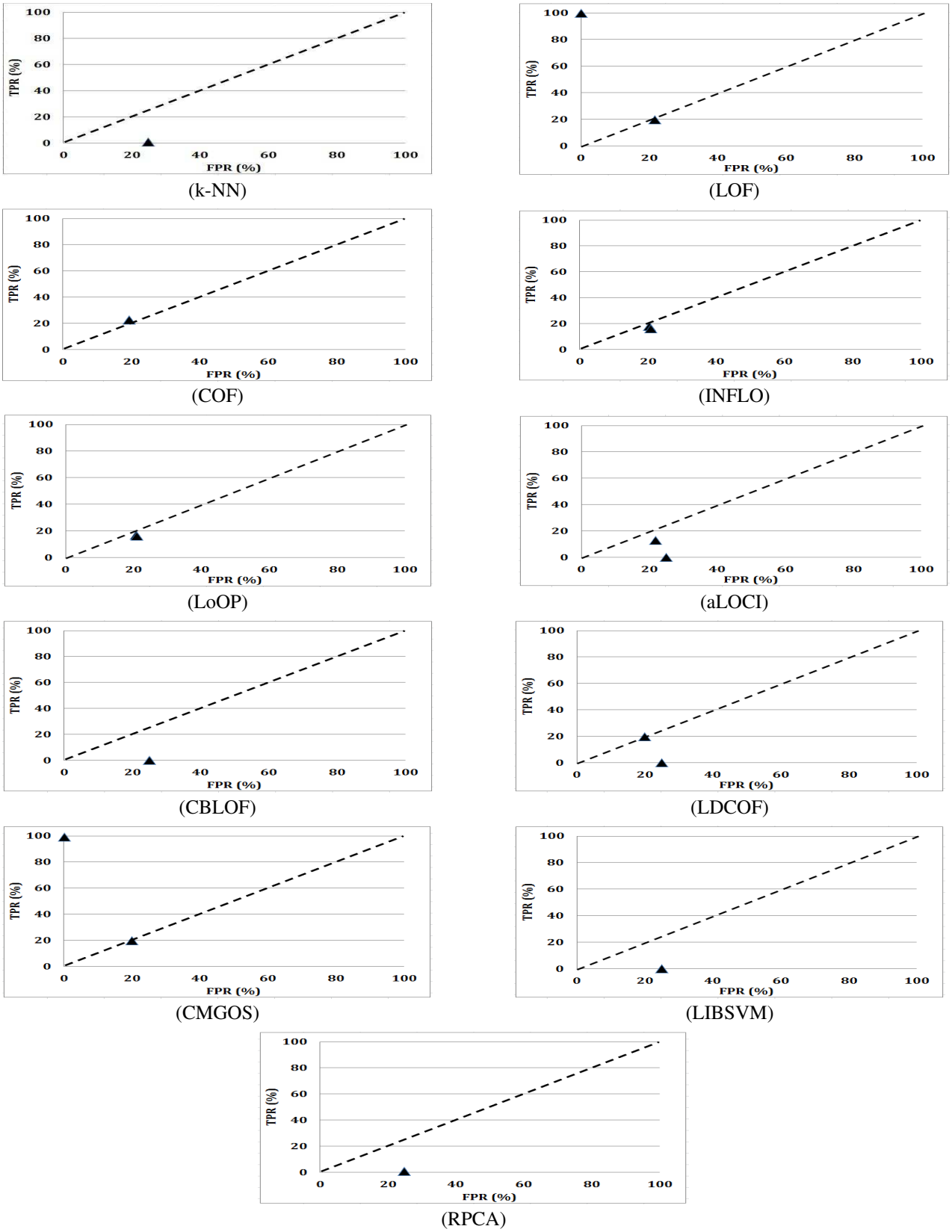
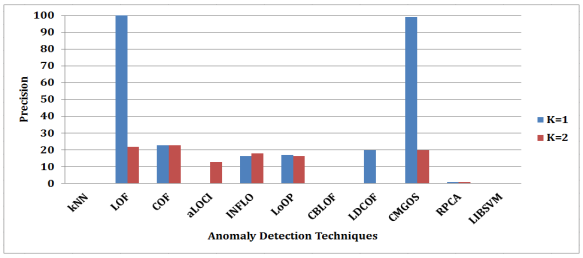


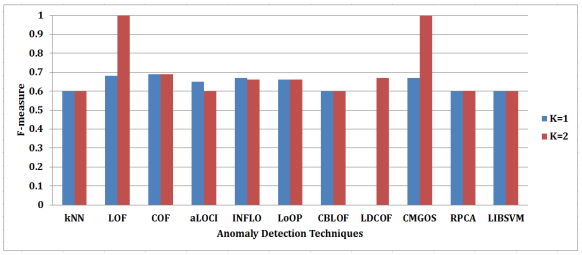
Fig. 1. True Positive vs False Positive Rate Comparison. The dashed line represents a random classifier.

TABLE I
BASIC STATISTICS OF TABLES STOCK DATASET USED IN THIS STUDY

Year	2009	2010	2011	2012	2013
Total number of stocks	16102	24144	23168	26350	16102
Average transaction days per stock	33	36	36	27	44
Average stocks per day	3052	3394	3351	2812	2838
Total volume of transactions (in billions)	808	933	990	703	691
Total value of transactions (in billions)	1445	1708	1678	1460	1642



(*Precision*)



(*F-measure*)

Fig. 2. Comparison using *Precison* and *F-measure*

transactions), in this study, we attempt to detect point/rare anomalies which is denoted by zero values for either of these five variables.

C. Key Insights from Experimental Results

For a robust validation of the anomaly detection techniques, we evaluate the performance using two different sets of samples from the original data maintaining the original ratio of normal and anomaly. For example, if the original data has a 3:1 ratio of normal and anomaly, the two different sample sets of data produced from the original data will also follow the same ratio. Therefore, the results will be free from performance bias when applied only on the single set of data. Fig. 1 shows the *FPR* and *TPR* performance of the anomaly detection methods. It is desirable for anomaly detection techniques to have high *TPR* and low *FPR*. The diagonal dashed line in the Fig. 1 corresponds to a Random Classifier. It is seen from the Fig. 1 that in all the cases the algorithms perform substantially poorer than the random classifier. In majority of the cases the *TPR* is close to 20% while the *FPR* is below 25% on average except in the case of *LOF* and *CMGOS* where the *TPR* is close to 100%. Fig. 2 reflects the *Precision* and *F-measure* comparison among the anomaly detection techniques where it is portrayed that the **LOF** and **CMGOS** are superior than the rest of the techniques.

IV. CONCLUSION

Unsupervised method of anomaly detection is normally used in exploratory settings where domain experts analyse the identified anomalies further to determine their application specific importance; however, in financial market, the amount of numerical data, upon which anomaly detection technique will work, is increasingly massive in size. Anomaly detection in big financial data can help the decision makers, like market participants, policy makers, and scholars, to mitigate the negative effects of misleading information. Therefore, in this study, following a detailed discussion on the popular unsupervised anomaly detection techniques, we attempted to apply them on financial market data. The objective was to analyze their performances in identifying rare anomalies. We came to a conclusion that nearest neighbour (**LOF**) and clustering based (**CMGOS**) approaches are more suitable for this domain than statistical and semi-supervised svm-based approaches. This study can further be extended on, how to incorporate the idea of collective and contextual anomaly in big data perspective and then how to introduce multi-view, hierarchical and co-clustering methods to improve the effectiveness of clustering-based anomaly detection techniques.

REFERENCES

[1] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137–144, 2015.
[2] T. Preis, H. S. Moat, and H. E. Stanley, "Quantifying trading behavior in financial markets using google trends," *Scientific reports*, vol. 3, p. srep01684, 2013.
[3] E. Yavrumyan, "Efficient market hypothesis and calendar effects: Evidence from the oslo stock exchange," Master's thesis, University of Oslo, 2015.
[4] M. Schulmerich, Y.-M. Leporcher, and C.-H. Eu, "Stock market anomalies," in *Applied Asset and Risk Management*. Springer, 2015, pp. 175–244.
[5] D. Turner, M. Schroeck, and R. Shockley, "Analytics: The real-world use of big data in financial services," *IBM Global Business Services*, pp. 1–12, 2013.
[6] D. Thesmar, J.-P. Bouchaud, P. Krueger, A. Landier *et al.*, "Sticky expectations and stock market anomalies," HEC Paris, Tech. Rep., 2016.
[7] M. Ahmed, A. Anwar, A. N. Mahmood, Z. Shah, and M. J. Maher, "An investigation of performance analysis of anomaly detection techniques for big data in scada systems," *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, vol. 15, no. 3, pp. 1–16, May 2015.
[8] M. Ahmed, A. N. Mahmood, and M. R. Islam, "A survey of anomaly detection techniques in financial domain," *Future Generation Computer Systems*, vol. 55, pp. 278 – 288, 2016.
[9] M. Ahmed, A. N. Mahmood, and J. Hu, *Outlier Detection*. New York, USA: CRC Press, January 2014, ch. 1, pp. 3–21, (in book: The State of the Art in Intrusion Prevention and Detection).
[10] W. Jin, A. K. Tung, J. Han, and W. Wang, "Ranking outliers using symmetric neighborhood relationship." in *PAKDD*, vol. 6. Springer, 2006, pp. 577–593.