Stock Price Manipulation Detection Using Empirical Mode Decomposition Based Kernel Density Estimation Clustering Method

Baqar Abbas(✉), Ammar Belatreche, and Ahmed Bouridane

Department of Computer and Information Sciences, Northumbria University, Newcastle, UK
baqar.rizvi@norhtumbria.ac.uk, {ammar.belatreche, ahmed.bouridane}@northumbria.ac.uk

Abstract. Stock market manipulation means illegitimate or illegal activities trying to influence the prices of stocks, hence diluting the legal definition of trading stocks. In this research, a model for detecting Stock price manipulation is presented for anomalies like Pump & Dump, Quote stuffing, Gouging or Spoof Trading. The model is presented on level 1-tick data which contains highly volatile time series and a high trading frequency making the detection more challenging. In literature, very less number of studies based on unsupervised learning for Stock market manipulation has been carried out. In addition, the existing studies focused only on specific anomalies and were not generalized enough to capture other anomalies. The research model used in this work uses unsupervised learning where the input data is decomposed using Empirical Mode Decomposition followed by Kernel Density Estimation based clustering technique for anomaly detection. One of the key advantages of this technique is receiver operating characteristic (ROC) curve, which is better than the currently available techniques and provides a maximum area under curve (AUC) equal to 0.96. The results in this work are also compared with existing benchmark approaches like K-means, Principal Component Analysis (PCA) based anomaly detection and Dirichlet process Gaussian Mixture Model (DPGMM) based anomaly detection, and a maximum improvement of 84% is obtained.

Keywords: Anomaly detection · Stock price manipulation · Empirical mode decomposition · Kernel density estimation · Clustering · K-means
Principal component analysis · Dirichlet process Gaussian mixture model

1 Introduction

The legal definition of trading stocks through financial markets becomes dilute when it involves some illegitimate actions trying to make profit. These illegitimate or illegal activities trying to influence the prices of stocks fall under the abstract definition of stock market manipulation. The sole intent of market manipulation is to create a deluded image of the price, which gives or is likely to give a false impression as to increase or decrease the demand or supply of the manipulated stocks. This ultimately leads to misleading interest in the manipulated stock, so to gain profit but using illicit

K. Arai et al. (Eds.): IntelliSys 2018, AISC 869, pp. 851–866, 2019.
https://doi.org/10.1007/978-3-030-01057-7_63

means [1]. Stock market manipulation has been categorized into three forms: Information based manipulation, Action based manipulation and Trade based manipulation [2]. Information based manipulation intends to spread a false rumor or release some inside information about a company or its stock with an intention to influence the price. Action based manipulation is an action rather than trading, performed by the company managers or executives who hold the supply of a well-established product by increasing its demand and hence the stock price. Four Kaupthing Bank executives were caught financing their own share purchases in large and hefty amounts arousing the interest of others [3]. Trade based manipulation on the other hand has everything to do with trading in a stock exchange where traders, investors or brokers buy/sell stocks by manipulating equity prices or volumes (number of shares or bonds etc. for any security, traded during a period of time) [2, 3]. The Securities and Exchange Commission (SEC), in a press release, 2015 charged Costa-Rica based MoneyLine Brokers Firm and its founder for engaging in "Pump & Dump" schemes to artificially inflate a stock's price of Warrior Girl, a former shell company and then sell their own shares [4]. According to the report, Moneyline and its subordinates made illegal profits estimated at a total of $2.3 million. One of the major types of trade-based manipulation is price manipulation in which the trader targets to influence the buy/sell prices of a financial security.

Market abusers perform fraudulent activities like spoof trading, pump and dump, ramping or gouging, etc. follow a sequence of well-defined actions or strategies to control the equity price of a stock. Pump and dump is a price manipulation scheme where the manipulator begins with a large volume of stocks purchased at a cheaper rate and then starts sending large amount of buy orders, creating a high demand of that security and hence the selling price of that security rises [5]. During this time, an impression is created among many investors about the increasing price of that specific stock which motivates them to add orders. When sufficient number of orders are added and the asked or selling price of that stock has been sufficiently increased the manipulator withdraws its bid orders and executes sell orders at the increased price at this moment. Another type of price manipulation tactic is ramping or gouging and momentum ignition also known as Spoof Trading [6]. For example, a manipulator wants to sell stock at a higher price than the current ask price. The manipulator will enter spoofed buy order in larger volume at a higher price than the current bid making other investors believe that this increased price is genuine expecting other legitimate investors to join. Once the order is matched, the manipulator will withdraw the large spoofing buy order, will then issue a sell order of large volume of shares, and would receive a sale at this manipulated price.

The main focus of this research is to detect the most common yet important type of trade-based manipulation called price manipulation. It should be noted that there are several challenges faced while detecting a manipulation in stock prices like; large and heterogeneous datasets; unlabeled datasets where abnormal and normal patterns are not marked in time series and the fact that there are different and evolving strategies of trade based manipulation [6]. In view of these challenges, unsupervised learning seems to be more pertinent approach to the matter. The major contribution of this research is as follows; this research recommends the combination of Empirical mode decomposition (EMD) followed by Kernel density estimation (KDE) based clustering for

anomaly detection for a selective set of features while examining two types of manipulation patterns. The rationale behind using EMD is the fact that it is a data-driven approach that does not require a priori the level of decomposition and also the basis function needed is extracted from analyzing the dataset, as it has to be specified in other decomposition methods [7]. KDE clustering helps in grouping the input data into clusters while fitting a Gaussian distribution without requiring the amount of clusters up front [8]. This makes it easier to analyze the data within a cluster because of its small size and better detection of price manipulation can be performed. The major advantage of using this approach is its decision-making capability based on analyzing the patterns that are subjected being an anomaly. Further, it also presents its distinctiveness from the existing benchmark approaches (unsupervised learning) for anomaly detection when comparing the Receiver Operating Characteristics (ROC) curve and the Area Under the Curve (AUC) for each ROC as demonstrated by the obtained experimental results. Another merit of applying this model is that it is not trained to a specific type of price manipulation scheme. That is, the detection is performed without any prior knowledge about the anomalies injected, be it their location in the time series or magnitude.

As the introduction of stock price manipulation provides a brief overview of the problem, Sect. 2 will highlight most of the research work carried out and explores the existing techniques that can be applied towards the detection of stock price manipulation. Further, Sect. 3 illustrates the flow of methodology implemented followed by data set used and then the analysis of the proposed research in the Experimental results upon comparison with the existing techniques as part of Sects. 4 and 5, respectively.

2 Related Work

A number of empirical studies have been conducted in detecting price manipulation but only a handful of them used unsupervised learning techniques. Ferdousi and Maeda [9] applied an unsupervised learning approach called peer group analysis to the stock manipulation and detected cases of manipulation with an appropriate level of success. However, they do not take into account the change of peer groups over time, which decreases the detection probability when some members in the same peer group may gradually exhibit distinct behavior from that of other members. Kim et al. [10] tried to improve peer group analysis approach by updating the size of the group with time but failed to identify the exact location in time of the suspicious activity. A market close ramping detection algorithm developed by Aitkens et al. [11]. An alert is detected if the difference between the closing price and the price fifteen minutes before exceeds a given threshold. The algorithm was able to detect when the threshold was set as the 99% histogram distribution cut off the historical price change during the corresponding time window. Palshikar et al. [12] proposed a method to detect stock price manipulation using collusion sets using graph clustering algorithms. It states that many manipulative cases in the stock market involved collusion sets. A collusion set is a group of traders who trade heavily among themselves. For this, instead of using real world dataset, they generated a synthetic database based on probability distributions

and collusion sets of different characteristics and sizes were injected. Furthermore, it also considered the whole dataset while analyzing rather than dividing it into smaller timestamps which will make the clustering process more robust. Islam et al. [13] tried to improve this work by considering purely circular collusion sets using Markov Clustering algorithm but did not address the similar problem of detection under timestamp. Cao et al. implemented [14] a semi-supervised learning approach toward price manipulation detection in stocks. The approach focused on decomposing the data using Dirichlet Process Gaussian Mixture Model (DPGMM) into different components defining normal and abnormal components and then trained a Markov model upon those components. Furthermore, the research had to specify the number of decomposition components, which is misleading as the distribution of the normal-abnormal patterns, might overlap with each other.

In this research, problem formulation is carried out by analyzing original stock price information, adding two types of synthetic anomalies that correlates with the existing manipulation activities in to it. Although the original data is injected with two types of anomalies per stock, a significant number of them are added at different time instants along the duration of the time series to check the robustness of its detection capability. Then, meaningful information in the form of innovative features that represents the manipulation pattern is extracted and finally a detection model for price manipulation is presented. Other than some of the methods employed in the literature, there are some existing methods towards anomaly detection using unsupervised learning techniques, namely, DPGMM based anomaly detection [15, 16], Principal Component Analysis (PCA) based anomaly detection [17] and K-means based anomaly detection. Following section reviews those methods by implementing them on the same dataset with added anomalies and compares their performance with the proposed approach.

2.1 Dirichlet Process Gaussian Mixture Model Based Anomaly Detection

A mixture model for making probability distribution is a mixture of multiple distributions. It is a weighted sum of multiple Gaussian distribution functions assigned to different subsets of data [18]. For, a given data set, x having d – dimensions it is assumed that it is drawn from a model having multiple Gaussian distributions. The data is grouped into K clusters as per different distributions whose means and variances are calculated using expectation maximization technique. Dirichlet Process is then employed followed by Gibb's sampling to calculate prior probability for each cluster's component parameters [18].

For a set of five different features, the data is first windowed with no overlapping between windows and then grouped into different clusters, their corresponding pdf learned. A threshold value is set that separates the normal and the anomaly regions in the pdfs of each component according to minimum data likelihood value adopted from the industry reference detection algorithm from Smart Group [11]: which has the 99% cumulative distribution cut-off. This means that data values falling in the region above 99.5% and below 0.5% value of cumulative probability are anomalies [14].

2.2 Principal Component Analysis Based Anomaly Detection

Principal Component Analysis is usually applied to reduce the number of dimensions of the input data set. It involves the transformation of a highly correlated input data into a set of components, orthogonal to each other. Among these, the first component having maximum variance or latent is the projection of the input data, having multiple dimensions, onto a single dimension. The data points are then further projected onto a new orthogonal dimension but having a lesser variance than the first component and the process is repeated until the stopping criteria is matched. An important property of principal components is that they are uncorrelated or are orthogonal to each other and the principal components are arranged in the order of decreasing variances [19]. Here the components having large variance are called major components and the ones smaller are called minor. This categorization is explained below:

Once the data set is projected onto several components: major and minor, anomaly detection approach is implemented. According to which, normalized major components, PCi (i = 1, 2…p) and minor components PCj (j = 1, 2…. q) are thresholded and categorized as follows:

$$\text{Anomaly if}, \begin{cases} \sum_{i=1}^{p} \frac{PC_i^2}{\lambda_i} > c_1 \in \text{major components} \\ \sum_{j=1}^{q} \frac{PC_i^2}{\lambda_j} > c_2 \in \text{minor components} \end{cases}$$

$$\text{Normal Instance if}, \begin{cases} \sum_{i=1}^{p} \frac{PC_i^2}{\lambda_i} \leq c_1 \in \text{major components} \\ \sum_{j=1}^{q} \frac{PC_i^2}{\lambda_j} \leq c_2 \in \text{minor components} \end{cases}$$

p number of major components
q number of minor components
λ_i Eigen values

Considering a window of the normalized components (major and minor) at a time (30 samples), the anomaly was detected. The divide between the number of major and minor components is that the top 50% of the variance for the original data set has the major components and the remaining 50% comprises of the minor components [20]. The value of c_1 and c_2 are decided heuristically in the approach [17], but 95% of the maximum value in each principal component is considered as a threshold here.

2.3 K – Means Clustering Based Anomaly Detection

K means is a process of grouping input data set into clusters with the nearest mean and variance [21]. Here, the data is partitioned into blocks or cells with its mean calculated using an iterative refinement similar to the expectation maximization approach in mixture models. The mean of a cluster so formed is also the centroid of the space defined for a given cluster. For the input data, a window of 100 observations is considered and passed on to the K means clustering. In order to detect anomalous data, once the clustering for the dataset is done, the intra cluster distance between each data

point and its centroid is calculated for every cluster using the Mahalanobis distance method. Mahalanobis distance method is used because of its utility in calculating the distance as per the transformation along the principal component axis in the cluster space [22]. Along with the intra-cluster distance so calculated, Mahalanobis distance between each cluster centroid and the points that are not clustered is also calculated. A threshold is applied on the distances so calculated and the data sample exceeding the threshold value is marked as an anomaly. The threshold used here is decided to be as the 90% of the maximum distance calculated within each cluster.

3 Methodology

The flow of the methodology used in this report is as follows, for a manipulated input time series containing stock prices, some artefacts restrain the detection of an anomaly (Fig. 1). In addition, as the time series is non-stationary in nature, its statistical properties like mean and variance for the high frequency components violently evolve with time and the distribution of prices deviates from normality. As the high frequency components of the time series are more prone to the anomalies, wavelet transform is employed to filter out the low frequency components in the signal i.e. only the high frequency components are considered and is used as a feature, $\hat{x}(t)$ [23] where $x(t)$ is the input time series (stock prices). The two anomalous patterns that describe the price manipulation are saw tooth and spike patterns as illustrated in Fig. 2a, b. The effect of such patterns needs to be captured in the features used. In order to do so, the stock price values, $x(t)$ and a new feature vector $w(t)$, that extracts only the change between two consecutive samples and then amplifies that difference if it exceeds a given threshold are selected as the feature values. Further, gradient of the price i.e. the rate of change of prices, the gradient of the new feature, $\frac{\partial(w(t))}{\partial t}$ that further magnifies the change are used as the feature sets.

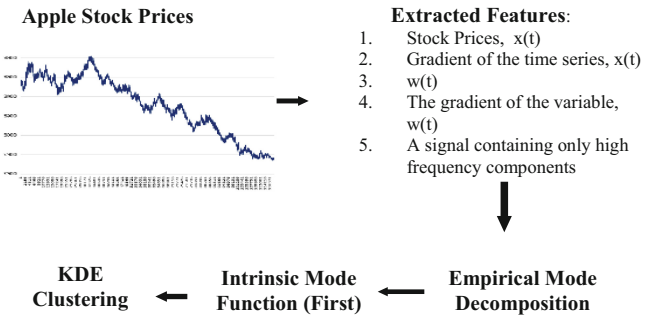

Fig. 1. Block diagram showing the approach.

Such a feature is calculated using discrete wavelet transform (DWT) where the input signal is decomposed up to single level into approximate and detail coefficients.

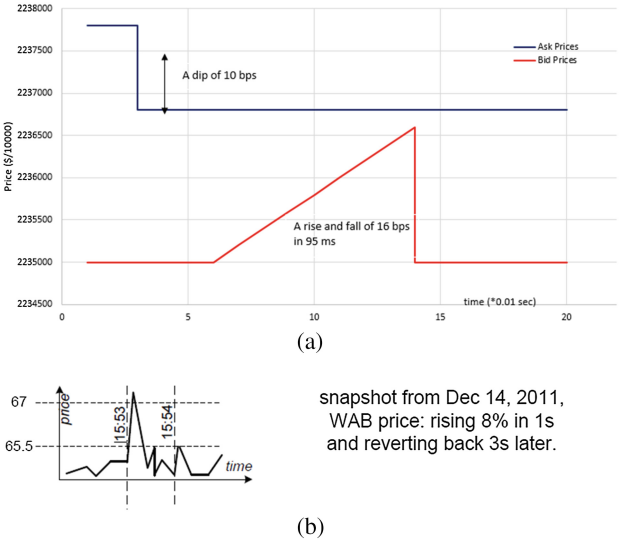

Fig. 2. Anomaly patterns (a) sawtooth (b) spike.

Approximate coefficients represent low frequency components and detail coefficients represent high frequency components.

$$X_{a,b} = \begin{cases} X_{a,b}, & X_{a,b} \leq \lambda \\ 0, & X_{a,b} > \lambda \end{cases} \tag{1}$$

A hard thresholding algorithm is then applied inversely on the detail coefficient, $X_{a,b}$ where a, b are shifting and scaling parameters for the given coefficient and λ is the threshold, so that the detail coefficients outside the threshold are set to zero. This threshold value in this case is calculated using universal threshold estimation method [24]. These filtered components are then reconstructed using Inverse DWT.

A feature set consisting of five individual features as follows:

(1) Input time series (Stock Prices), $x(t)$
(2) Gradient of the price time series, $\frac{\partial(x(t))}{\partial t}$
(3) A new univariate variable feature $w(t)$ explained below.
(4) The gradient of the new variable $\frac{\partial(w(t))}{\partial t}$ and
(5) A signal containing only high frequency components, $\hat{x}(t)$ were considered.

The feature set $w(t)$ is described as follows:

$$s(t) = x(t) - x(t-1) \tag{2}$$

$$w(t) = \begin{cases} 3 * s(t), & s(t) > threshold \\ s(t), & s(t) \leq threshold \end{cases} \tag{3}$$

where $x(t)$ is the input time series and $s(t)$ is the difference between two consecutive samples. Typically, a threshold value of 3 bps is selected.

3.1 Empirical Mode Decomposition (EMD)

EMD is a process of decomposing a time series into components that preserves the characteristics of the varying frequency as that of the original signal and are called intrinsic mode functions (IMFs). These decomposed components are orthogonal to each other and to the original signal, are of the same length as that of the original signal and remains in the time-domain [25]. Since the decomposition is based on the analysis of local time scale of the data and since the obtained components (IMFs) provides instantaneous frequencies as functions of time, it can be applied to non-linear and non-stationary process. An IMF has same number of maxima and minima throughout the duration of the signal and the mean value within an envelope having maxima and minima will be zero i.e. it will have equal number of positive and negative values within a localized envelope. The process of calculating an IMF is called the sifting process. According to which, first the mean (m_1) of the upper and lower envelope of the original signal is calculated using cubic-spline interpolation method [26]. The difference between x(t) and m_1 is the first component (4a), which should ideally satisfy the conditions for IMF.

$$x(t) - m_1 = s_1 \tag{4a}$$

However, if it does not, the process is repeated now considering the difference as the new signal and further calculations of upper and lower envelope's mean unless the new difference satisfies the condition of being an IMF.

$$s_1 - m_{11} = s_2 \tag{4b}$$

An IMF, so calculated will be the first IMF component, $r(t)$ of the original time series,

$$s_k - m_{1k} = r(t) \tag{4c}$$

Then, the first obtained IMF is separated from the original signal,

$$x(t) - r(t) = x_1(t) \tag{4d}$$

This process is again repeated for $x_j(t)$, until the number of zero-crossings and the number of extrema is the same or almost differ by one. A situation in which the

resulting signal becomes mono-component i.e. it has no negative frequency component [25]. The first IMF contains most of the high frequency components, which can be considered as random noise, but is the most interesting feature while tracking down anomaly-effected portions of the signal (high frequency) [27]. So, for all of the proposed five features, one dimensional empirical mode decomposition is applied to each feature and the first IMF of each feature so calculated is preserved and the rest are forsaken. The IMF values will now act as an input to the clustering algorithm via Kernel Density Estimation (KDE) approach.

3.2 KDE Clustering Based Anomaly Detection

KDE clustering based anomaly detection is a modified approach for anomaly detection via non-parametric density estimation for clustering. It has the advantage that it does not require a prior knowledge of the number of clusters. The method suggests calculating a kernel based density estimation for a set of data samples and cluster them based on the following algorithm [8]. For an input data sample 'x',

$$x = \{x_1, x_2, x_3 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots x_n\}$$

The kernel density estimator used to calculate the probability density $\hat{f}(x)$ is given by,

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{5}$$

where, 'n' is the length of the data to be clustered, X_i is the standard deviation of the data 'x', 'h' is the sample size or bandwidth or the window, the kernel function, K that is Gaussian here,

$$K(x) = \frac{1}{2\pi} exp\left(-\frac{x^2}{2}\right) \tag{6}$$

Given: $x = \{x_1, x_2, x_3 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots x_n\}$ be a univariate vector that is to be clustered and $\alpha \in R$; the bandwidth h is defined as follows,

$$h = 1.06\sigma n^{-1}\alpha;\ for\ Gaussian\ Kernel \tag{7}$$

Considering the length of the cluster C to be zero at first, the algorithm suggests that given the original input data set, X, bandwidth parameter, h is calculated (7). Based on which, if the difference between the mean of the sample points (x) calculated and the sample points is less than 'h' are grouped into one cluster, C_1. Now, the length of the input vector is reduced by the number of data points already clustered. For the remaining data, bandwidth parameter (h) and standard deviation, X_i' are again calculated and the same process continues until all the points in the input vector are clustered into 'j' clusters. 'α' is a parameter calculated for the kernel density estimation and

whose value is set to 5 as proposed by Silverman [28]. 'σ' is the standard deviation of X, 'n' is the length of X that keeps on changing at every iteration. Now, within each cluster so formed, each cluster has a different distribution as shown in Fig. 3a, b. The values on the horizontal axis are the random values taken over by the feature set w(t) and on the vertical axis, probability density. For each cluster, feature samples having a probability density, calculated in (5), less than 0.5% is marked as an anomaly. In this way EMD based KDE clustering approach suitably identifies the exact location in time, when the manipulation occurred and provides better performance compared with the literature.

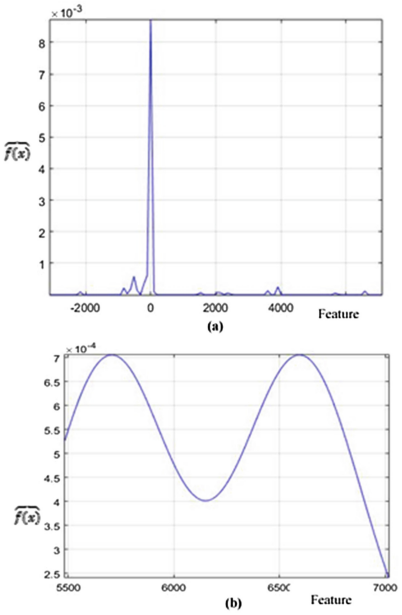

Fig. 3. Probability distribution for clusters (a) & (b)

A comparison of the results so calculated with some of the existing approaches for unsupervised learning in anomaly detection like Dirichlet process Gaussian Mixture Model, K-Means, Principal Component Analysis is shown in the next sections. It should be noted that none of these approaches had been used for stock market manipulation in the literature. The next section explains the dataset used and details of how these existing approaches are implemented considering above mentioned features as the input data set for them.

4 Dataset Used

A level 1 – tick dataset of five major stock companies viz. Apple, Amazon, Google, Microsoft and Intel Corp. (NASDAQ Stock Exchange, USA) are selected from LOBSTER project [29] because of their high volatility, high trading frequency which makes such stocks more prone to manipulation [5, 30] and higher trading volumes. Buy/Sell orders or Bid/Ask, as it is known by these names in an exchange of any securities like those that stock shares, bonds etc. are executed through an order book. An order book is a business list of buy and sell orders for a stock, share or any other financial instrument, organized as per the prices. An order book also lists the volume of shares or stocks being bid or asked at a certain price level also known as depth of the order book [31]. The size of each dataset varies per stock Apple, Amazon and Google roughly converge themselves within 200,000 samples, whereas the rest of the stock like Intel Corp and Microsoft have larger sizes of 800,000 samples. Based on this, they are categorized into two groups (Group I & II). Group I consists of Apple, Amazon, and Google stocks and Group II has Intel Corp and Microsoft Stock prices.

5 Experimental Results

The data set used in this paper is taken from an open source LOBSTER database [29] consisting of Apple, Amazon, Google, Intel Corp and Microsoft stocks from 12th June, 2012. Each stock has a level 1-tick data and the number of samples varies with different stocks. Based on it, the data set is divided into two groups (Group I & II). Each stock data is reportedly having no manipulation of any sort [32]. *Group I* has three stocks: Apple, Amazon and Google having around 200,000 samples and *Group II*: Intel Corp and Microsoft' Stocks consisting of more than 800,000 sample data points.

An artificial anomalous database is generated in order to test the validity and robustness of the proposed approach. As explained in Sect. 1, there are two types of anomalies injected in the original data samples. Type 1 is a synthetic anomalous waveform having a saw-tooth like fall of 16 bps in 95 ms and Type 2 have a rise and then sudden fall of 30 bps in a time span of 0.1 s as shown in Fig. 2a, b. These anomalies are then injected into the corresponding original time series making it a mixture of both normal and anomalous waveforms. To ensure comprehensive assessment of the approach, *group I* is injected with 50 anomalies of each type, making a total of 100 anomalies in them and *group II's* stocks are injected with 200 anomalies of each type making a total of 400 anomalies in it. The place of injection for an anomaly in a time series is performed without taking into account of the time and preceding and succeeding information of the price to make the anomaly detection more challenging. It is even possible that it may be directly followed by a similar waveform but any prior knowledge of the data set is totally avoided.

Or all of the proposed and existing approaches, their performance is evaluated using Receiver Operating Characteristics (ROC) curve. In order to calculate the ROC, some of its parameters need to be explained first. (i) True Positives represent the total number of normal instances correctly detected as normal, (ii) True Negatives represent the total number of anomalous samples correctly detected as anomalies, (iii) False

Positives represent the total number of anomalies incorrectly detected as normal instance, and (iv) False Negatives represent the total number of normal samples incorrectly detected as anomalies [33]. In this paper, ROC curve is plotted between True Positive Rate (TPR) and False Positive Rate (FPR), where TPR = TP/(TP + FN) and FPR = FP/(FP + TN) are calculated while varying window size of the input data for KDE clustering.

The ROC curves for five stocks with 200 and 400 number of anomalies injected are shown in Fig. 4. The tweaking factor in the calculation of a varying TPRs and FPRs is the threshold applied on the output score for each approach for anomaly detection ranging [0, 1]. The output score for different approaches are as follows; cumulative probabilities for DPGMM based approach, normalised Principal components values for PCA based approach, normalised distance measure for K-Means based approach and cumulative probabilities obtained from kernel density estimates. The summary of the AUC values for different techniques used are condensed in the form of a comparator chart in Fig. 5.

The AUC values for EMD based approach and its dominance over other existing approaches clearly indicates the better performance for all the five stocks. As from Fig. 5, it can be shown that the proposed approach retains their advantage in terms of anomaly detection over the existing approaches and can achieve relatively higher values for AUC. As an example, the approach using raw features as an input to the KDE clustering algorithm maintains a stability in achieving high TPR for almost all of the stocks while K-Means, PCA and DPGMM didn't perform well in some of the stocks and remained volatile.

6 Discussion

The experimental results obtained by the EMD – KDE based approach have shown a significant development in achieving a higher rate of detection of manipulation of two types (Saw tooth & Spike pattern) in the price. These manipulative actions relating to pump & dump, ramping and quote stuffing are carefully selected as they seem to provide similar impact on price of a stock as the ones depicted by these added anomalies. The results also outperformed some of the existing approaches using unsupervised learning for anomaly detection. The robustness of the proposed approach can be explained from the decomposition of the feature sets for a given length of the samples in a window using EMD. The fact that, while considering the cases of price manipulation for pump & dump and quote stuffing, a sudden flip in the prices happens after a long held position of incremental rise in prices. So, the window size for the decomposition of the dataset should be carefully selected taking into account only two components for a given window, one that explains the constant positive or negative slope and the other that represents the sudden drop. Another reason that contributes to its robustness is the threshold value that needs to be set up on such components or the margin that can probably divide the normal and anomalous boundary.

The EMD based approach followed by KDE clustering for manipulation detection achieved the highest AUC on all of the five stocks and outperformed all of the existing models for unsupervised learning towards anomaly detection. The best AUC is

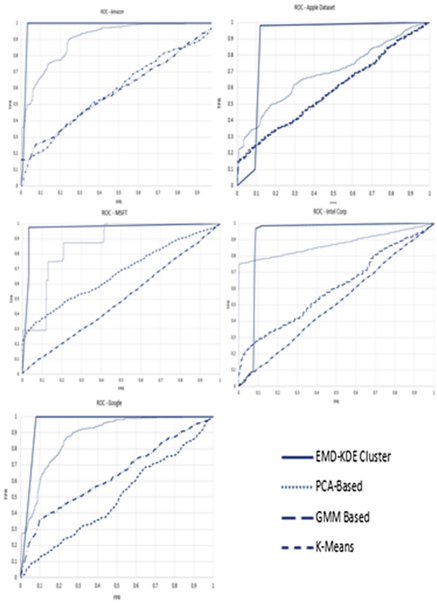

Fig. 4. ROCs of the five stocks for five different models.

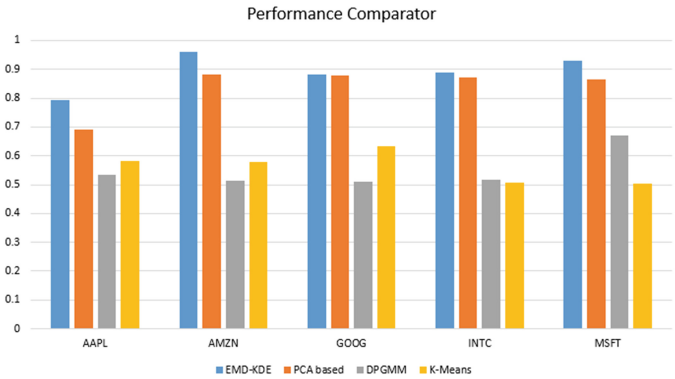

Fig. 5. Performance comparison based on AUC.

achieved for the Amazon stocks (0.9623) which is about 6.7% higher than the PCA based approach, 65% higher than the K-Means based approach and almost double than under approaches using Dirichlet process GMM and using only raw features. The second best detection is for Microsoft stocks (0.9308) which is 7.5, 84 and 38.6% higher than using the PCA based approach, the K-Means based detection and the DPGMM based detection, respectively. The lowest performance for the proposed approach is observed with Apple stocks (0.7946) which is still higher by 15%, than the

PCA based approach and 36% higher than the K-Means based approach and 48% higher than the DPGMM based approach. A performance comparison graph based on the values of AUC for all the stocks and for each different approach applied is shown in Fig. 5.

The EMD – KDE clustering based approach for manipulation detection performed variably for some of the data sets and did not attain very high values of AUC as it did for Amazon and Google stocks. This can be attributed to the high variability of the data and the mixing of the anomalies with similar waveforms that created large False Positives (FPs) but still it managed to get AUC values higher than the rest of the existing approaches.

7 Conclusion

This paper presented an innovative approach for detecting stock price manipulation based on EMD and KDE clustering. A brief review of the literature covering detection of market abuses has also been presented along with their limitations. This research envisages two types of manipulations existing in the stock markets, which relates to different categories of price manipulation and strives to work upon their detection using unsupervised learning. To achieve this, a large open source database, which is known for not having any manipulation, is considered. To test the validity of the proposed approach, a very large number of artificially generated anomalies are then injected to it making the input dataset, a mixture of both normal and manipulated instances. Based on the extracted features, instantaneous mode functions (IMFs) were computed using the EMD algorithm. Once IMFs are obtained for a given stock, the dataset is then windowed before passing these to the KDE clustering algorithm for manipulation detection.

KDE clustering algorithm groups the input data set, based on the density estimate defined within a bandwidth parameter, into clusters. A threshold value set up on the value of the pdf for a given cluster separates the normal and anomalous samples. It is found that the proposed model outperforms the existing approaches by a maximum of 84% higher than the AUC for some stocks.

References

1. Financial Conduct Authority: MAR 1.6 market abuse, sec. 118, no. 5 (2012)
2. Allen, F., Gale, D.: Stock price manipulation. Rev. Financ. Stud. **5**(3), 503–529 (1992)
3. Securities and Exchange Commission 2015: Case: 1:15-cv-05456. Available from: https://www.sec.gov/news/pressrelease/2015-146.html
4. Wade, R.H., Sigurgeirsdottir, S.: Iceland's meltdown: the rise and fall of international banking in the North Atlantic. J. Rev. Economia Polit. São Paulo **31**(5), 684–697 (2011)
5. Neupane, S., Rhee, S.G., Vithanage, K., Veeraraghavan, M.: Trade-based manipulation: beyond the prosecuted cases. J. Corp. Finan. **42**, 115–130 (2017)
6. Lee, E.J., Eom, K.S., Park, K.S.: Microstructure-based manipulation: strategic behavior and performance of spoofing traders. J. Financ. Markets **16**(2), 227–252 (2013)

7. Labate, D., Foresta, F.L., Occhiuto, G., Morabito, F.C., Ekuakille, A.L., Vergallo, P.: Empirical mode decomposition vs. wavelet decomposition for the extraction of respiratory signal from single-channel ECG: a comparison. IEEE Sens. J. **13**(7), 2666–2674 (2013)
8. Matioli, L.C., Santos, S.R., Kleina, M., Leite, E.A.: A new algorithm for clustering based on kernel density estimation. J. Appl. Stat. **44**(1), 1–20 (2016)
9. Ferdousi, Z., Maeda, A.: Unsupervised outlier detection in time series data. In: Proceedings of the 22nd International Conference on Data Engineering Workshops, pp. 51–56 (2006)
10. Kim, Y., Sohn, S.Y.: Stock fraud detection using peer group analysis. Expert Syst. Appl. **39** (10), 8986–8992 (2012)
11. Aitken, M.J., Harris, F.H., Ji, S.: Trade-based manipulation and market efficiency: a cross-market comparison. In: 22nd Australasian Finance and Banking Conference, Sydney, pp. 1–43 (2009)
12. Palshikar, G.K., Apte, M.M.: Collusion set detection using graph clustering. Data Min. Knowl. Disc. **16**(2), 135–164 (2007)
13. Islam, M.N., Haque, S.M.R., Alam, K.M., Tarikuzzaman, M.: An approach to improve collusion set detection using MCl algorithm. In: 12th International Conference on Computers and Information Technology, 2009, ICCIT '09, Dhaka, pp. 237–242 (2009)
14. Cao, Y., Li, Y., Coleman, S., Belatreche, A., McGinnity, T.M.: Adaptive hidden Markov model with anomaly states for price manipulation detection. IEEE Trans. Neural Netw. Learn. Syst. **26**(2), 318–330 (2015)
15. Yeung, D.Y., Ding, Y.: Host-based intrusion detection using dynamic and static behavioral models. Pattern Recognit. **36**(1), 229–243 (2003)
16. Clifton, D.A., Tarassenko, L., Sage, C., Sundaram, S.: Condition monitoring of manufacturing processes. In: Proceedings of Condition Monitor, pp. 273–279 (2008)
17. Shyu, M.L., Chen, S.C., Sarinnapakorn, K., Chang, L.W.: A novel anomaly detection scheme using principal component classifier. In: Proceedings of the 3rd IEEE International Conference on Data Mining, pp. 353–365 (2003)
18. Neal, R.M.: Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Stat. **9**(2), 249–265 (2000)
19. Jolliffe, I.: Principal Component Analysis. Wiley Online Library (2002)
20. Abdi, H., Williams, L.J.: Principal component analysis. Wiley Interdisc. Rev. Comput. Stat. **2**(4), 433–459 (2010)
21. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
22. Li, S.Z., Jain, A. (eds.): Mahalanobis distance. In: Encyclopedia of Biometrics. Springer US, Boston, MA, p. 953 (2009)
23. Haven, E., Liu, X., Shen, L.: De-noising option prices with the wavelet method. Eur. J. Oper. Res. **222**(1), 104–112 (2012)
24. Donoho, D.L., Johnstone, I.M.: Ideal Spatial Adaptation by Wavelet Shrinkage. Department of Statistics, Stanford University, USA (1993)
25. Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, E.H., Zheng, Q., Tung, C.C., Liu, H.H.: The empirical mode decomposition method and the Hilbert spectrum for non-stationary time series analysis. Proc. R. Soc. Lond. **454**, 903–995 (1998)
26. Mandic, D.P., Rehman, N., Wu, Z., Huang, N.E.: Empirical mode decomposition-based time-frequency analysis of multivariate signals. IEEE Sig. Process. Mag. **74**, 74–86 (2013)
27. Hong, L.: Decomposition and forecast for financial time series with high-frequency based on empirical mode decomposition. Energy Procedia **5**, 1333–1340 (2011)
28. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapman and Hall, London (1986)
29. LOBSTER project, Atlanta, GA, USA. Limit order book system [Online] (2012). Available: http://www.lobster.wiwi.hu-berlin.de

30. Cumming, D.J., Zhan, F., Aitken, M.J.: High frequency trading and end-of-day manipulation. York University, Toronto, ON, Canada, Technical Report, pp. 1–34 (2013)
31. Cumming, D., Johan, S., Li, D.: Exchange trading rules and stock market liquidity. J. Financ. Econ. **99**(3), 651–671 (2011)
32. Tse, J., Lin, X., Vincent, D.: High frequency trading—measurement, detection and response. Credit Suisse, Zürich, Switzerland, Technical Report (2012)
33. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)