Time Series Contextual Anomaly Detection for Detecting Market Manipulation in Stock Market

Koosha Golmohammadi, Osmar R. Zaiane
Department of Computing Science
University of Alberta
Edmonton, Canada
{golmoham, zaiane}@ualberta.ca

Anomaly detection in time series is one of the fundamental issues in data mining that addresses various problems in different domains such as intrusion detection in computer networks, irregularity detection in healthcare sensory data and fraud detection in insurance or securities. Although, there has been extensive work on anomaly detection, majority of the techniques look for individual objects that are different from normal objects but do not take the temporal aspect of data into consideration. We are particularly interested in contextual outlier detection methods for time series that are applicable to fraud detection in securities. This has significant impacts on national and international securities markets. In this paper, we propose a prediction-based Contextual Anomaly Detection (CAD) method for complex time series that are not described through deterministic models. The proposed method improves the recall from 7% to 33% compared to kNN and Random Walk without compromising the precision.

Keywords: *Anomaly detection, outlier detection, data mining, financial time series, fraud detection*

I. INTRODUCTION

Anomalies or outliers are individuals that behave in an unexpected way or feature abnormal properties [1]. The problem of identifying these data points or patterns is referred to as outlier/anomaly detection. The significance of anomaly detection lies in actionable information that they provide in different domains such as anomalous traffic patterns in a computer networks which may represent intrusion [2], anomalous MRI images which may indicate the presence of malignant tumors [3] anomalies in credit card transaction data which may indicate credit card or identity theft [4], or anomalies in stock market which may indicate market manipulation. Detecting anomalies has been studied by several research communities to address issues in different application domains [5]. Time series are indispensible in today's world and collected data in many domains such as computer network traffic, healthcare, flight safety, fraud detection etc. are sequences or time series. More formally, a time series $\{x_t, t \in T_0\}$ is the realization of a stochastic process $\{X_t, t \in T_0\}$. For our purposes the set T (i.e. set of time points) is a discrete set and the real valued observations x_t are recorded on fixed time intervals. Although, there has been extensive work on anomaly detection [5], majority of the techniques look for individual objects that are different from normal objects but do not take the temporal aspect of data into consideration. For example, a conventional anomaly detection approach based on values of data points may not capture anomalous data points in the ECG data in Figure 1. Therefore, the temporal aspect of data should be considered in addition to the amplitude and magnitude values.

Although, time series anomaly detection methods constitute a smaller portion of the body of work in anomaly detection, there has been many methods within this group that are designed for different domains. Time series outlier detection methods are successfully applied to different domains include management [6], detecting abnormal conditions in ECG data [7], detecting

Figure 1. Anomaly in ECG data (representing second degree heart block)

shape anomalies [8], detecting outlier light curves in astronomical data [9] and credit card fraud detection [10]. These methods are shown to be effective in their target domain but adapting the methods to apply to other domains is very challenging. This is evidently due to the fact that the nature of time series and anomalies are fundamentally divergent in different domains. We are particularly interested in developing effective outlier detection methods for complex time series that are applicable to fraud detection in securities (stock market). The significance of detecting such outliers is due to the fact that these outliers by definition represent unexpected (suspicious) periods which merit further investigations as they are potentially associated to market manipulation.

A. Problem Setting

The outlier detection problem for time series data can be perceived in three settings:

1. Detecting anomalous time series, given a time series database: here the time series is anomalous with respect to the training time series in the database. The time series in the database may be labeled or a combination of labeled and unlabeled samples.
2. Detecting anomalous subsequence: here the goal is identifying an anomalous subsequence within a given long time series (sequence). This problem setting is also introduced in the works of Keogh et. al. as detecting discords – "the subsequences of a longer time series that are maximally different from the rest of the sequence" [11] – in

time series [12]. Figure 2 shows an anomalous subsequence within a longer time series. It is not the low values of the subsequence which make it anomalous, as it appears in other places in the given time series, but it is abnormal length of the subsequence.

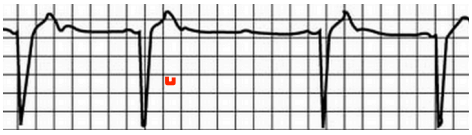

Figure 2. Anomalous subsequence within a longer time series

3. Detecting contextual or local outliers: here anomalies are data points that are anomalous in a "specific context but not otherwise". For example, the average temperature of Edmonton during 2013 (see Figure 3) is 4.03 degrees Celsius while the same value during January would be an outlier (i.e. contextual outlier). Another example would be data points or periods in a time series that deviate from the expected pattern given a group of time series that are expected to have a similar pattern (e.g. heart rate of different horses or stock returns of similar companies).

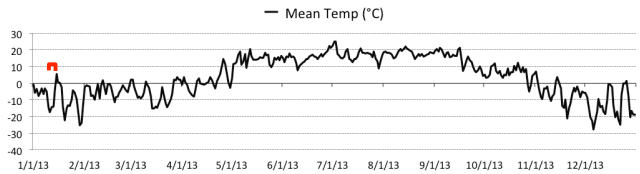

Figure 3. Average daily temperature of Edmonton during the year 2013

In this paper we focus on contextual/local anomaly detection within a group of similar time series. The context is defined both in terms of similarity to the neighborhood data points of each time series and similarity of time series pattern with respect to the rest of time series in the group. Local anomalies/outliers are different from global outliers because a data point that is detected as an outlier with respect to the neighborhood data points may not be an outlier with respect to all other data points in the dataset. Local outlier detection methods are particularly useful in non-homogenous datasets and datasets with changing underlying factors such as financial data. The major motivation of studying local outlier detection for us is developing methods for detecting local outliers in complex time series that do not follow a seasonal pattern and are non-parametric, meaning it is difficult to fit a polynomial or deterministic function to the time series data. This is a significant problem in domains with complex time series such as stock market. It has been shown that market manipulation periods are associated with outliers in the time series of assets [13] [14], yet the problem of developing effective methods to detect such outliers remains a challenging problem.

In Section 2, we discuss challenges in developing outlier detection methods for time series, review existing methods for detecting contextual outliers, and highlight drawbacks of these methods in comparison to our proposed method. In Section 3, we introduce the proposed method and details of implementation. In Section 4, we discuss experimental results.

II. RELATED WORKS

We reviewed the literature on different data mining methods for detecting securities market manipulation in an earlier work [15]. In this section, we focus on characteristics and drawbacks of existing methods. We elaborate on our approach towards addressing limitations of the existing methods.

Anomaly detection methods for detecting contextual outliers in time series can be classified along two orthogonal directions: i) the way the data is transformed prior to anomaly detection (transformation dimension), and ii) the process of identifying anomalies (anomaly detection technique). Table 1 describes a list of existing methods for detecting local outliers in time series along these two dimensions.

Transformation is the procedure that is applied to data before anomaly detection. There are two motivations for data transformation: i) to handle high dimensionality, scaling and noise, and ii) to achieve computational efficiency. The transformation procedures include:

- **Aggregation** that focuses on dimensionality reduction by aggregating consecutive values. A typical approach for aggregation is replacing a set of consecutive values by a representative value of them (usually their average).
  Discretization which converts the given time series into a discrete sequence of finite alphabets. The motivation of using discretization is using existing symbolic sequence anomaly detection algorithms and improving computation efficiency [16].
- **Signal Processing** which maps the data to a different space as sometimes detecting outliers in a different space is easier and the mapping may reduce the dimensionality (e.g. Fourier transforms [17], wavelet transforms [18]).

There are some issues and risks that need to be considered when using transformation techniques. The time series are in a different format after aggregation, therefore, the values after transformation correspond to a set of data points in the original time series. This is particularly problematic in time series that do not follow a uniform distribution. Although, discretization may improve computational efficiency, but the dimensionality of symbolic representations remains the same after transformation. Most discretization techniques need to use the entire time series to create the alphabet. Furthermore, the distance measures on symbolic representation may not represent a meaningful distance in the original time series. Transformation using signal processing techniques may also suffer from the issue of distance measure in the new space. We avoid the transformation process in the proposed outlier detection method and we use original values of all data points in the given time series. The time series in securities fraud detection are typically processed offline (there is no noise in recorded values) and are aligned time series.

Below, we briefly review five groups of anomaly detection methods to detect local/contextual outliers in time series and we highlight their disadvantages:

- **Window based:** a time series is divided to fixed window size subsequences. An anomaly score is calculated by measuring the

Table 1. Anomaly Detection Methods for Time Series

Transformation → / ↓ Technique	Aggregation	Discretization	Signal Processing
Window Based	kNN [19], SVM [20]	kNN [19]	
Proximity Based	PCAD [21], [22]		
Prediction Based	Moving Average [23], AutoRegression [23], Kalman Filters [24], SVM [25]	FSA [26]	Wavelet [18] [27]
HMM based	[28]	[29] [30]	
Segmentation	[31] [32] [33]		

distance of a sliding window with the windows in the training database. Chandola et al. use the distance of a window to its k^{th} nearest neighbor as the anomaly score [19] while Ma et al. use the training windows to build one class SVMs for classification (the anomaly score for a test window is 0 if classified as normal and 1 if classified as anomalous) [20].

- Disadvantage: the window based outlier detection methods for time series suffer from two issues: i) the window size has to be chosen carefully (the optimal size depends on the length of anomalous subsequence), and ii) the process can become computationally expensive (i.e. $O((nl)^2)$ where n is the number of samples in testing and training datasets and l is the average length of the time series).

In our proposed method, we divide the given time series to fixed window size periods and look for outliers within that period (i.e. neighborhood) but there is no sliding window (thus lower time complexity). Furthermore, the size of windows in the proposed method (e.g. 1 years) is much longer than the length of anomalies. We use overlapping of a few time stamps to avoid missing outliers on the border of the windows. The length of the overlapping is set to 4 data points in our experiments.

∞ **Proximity based:** the assumption here is that the anomalous time series are different to other time series. These methods use the pairwise proximity between the test and training time series using an appropriate distance/similarity kernel (e.g. correlation, Euclidean, cosine, DTW measures). Unlike the window based method, instead of rolling a window the similarity measure is used to measure the distance of every two given sequences. A k-NN or clustering method (k-means) is used where the anomaly score of each time series is the distance to the k^{th} nearest neighbor in the dataset in the former case, and the distance to the centroid of the closest cluster in the latter case [34] [21].

- Disadvantage: these methods can identify anomalous time series, but cannot exactly locate the anomalous region. They are also highly affected by the similarity measure that is used, and in the problems that include time series misalignment the computational complexity may significantly increase.

Our proposed method, like any outlier detection method which uses a distance measure, is affected by the type of distance measure, however, it has been shown that the Euclidean distance (the similarity measure that we use) outperforms most distance measures for time series [35]. As we indicated in Section 1, the time series in our problem are discrete and the values are recorded in fixed time intervals (i.e. time series are aligned). Unlike proximity based methods, which assign an anomaly score based on the distance of two given sequences, our proposed method assigns an anomaly score based on the distance of predicted value for each data point and its actual value, thus enables detecting the location of anomalous data point/region.

∞ **Prediction based:** these methods assume the normal time series is generated from a statistical process but the anomalous data points do not fit the process. The time series based models such as Moving Average (MA) [23] Auto Regressive (AR) [23], Autoregressive Integrated Moving Average (ARIMA) [36] and ARMA [37] as well as non-time series based models such as linear regression [38], Gaussian process regression [39] and support vector regression [25] are used to learn the parameters of the process. Then, the model derived from a given time series is used to predict the $(n+1)^{th}$ value using previous n observations.

- Disadvantage: there are two issues in using such prediction based methods for outlier detection in time series: i) the length of history that is used for prediction is critical in locating outliers, and ii) performance of these methods are very poor in capturing outliers if the data is not generated by a statistical process.

The assumption that the normal behavior of any given time series is generated from a model and such a model could be derived from history of the time series, does not hold in some domains such as securities market[1]. Therefore, outlier detection methods based on this assumption (i.e. prediction based, Hidden Markov Model and segmentation based) are inappropriate in detecting anomalies in complex time series such as securities.

∞ **Hidden Markov Model (HMM) based:** the assumption here, is the underlying process creating the time series is a hidden Markovian process (i.e. the observed process creating the original time series is not necessarily Markovian) and the

[1] If such a model could be devised, one would be able to predict

normal time series can be modeled using an HMM [40] [41]. The training data is used to build an HMM which probabilistically assigns an anomaly score to a given test time series.

- Disadvantage: the issue in using HMM based methods is the assumption that there is a hidden Markovian process generating the normal time series. Therefore, this method fails if such a process does not exist.

∞ **Segmentation based:** first a given time series is partitioned into segments. The assumption here is that there is an underlying Finite State Automaton (FSA) that models the normal time series (the states and transitions between them in FSA is constructed using the training data) and segments of an anomalous time series do not fit the FSA [31] [32].

- Disadvantage: segmentation based methods may suffer from two issues: i) the state boundaries are rigid and may not be robust to slight variations in the data during the testing phase, and ii) segmentation technique may fail in detecting outliers in problems where the assumption "all training time series can be partitioned into a group of homogeneous segments" does not hold.

III. METHOD

The classic approach in anomaly detection is comparing the distance of given samples with a set of normal samples and assigning an anomaly score to the sample. Then, samples with significant anomaly scores are labeled as outliers/anomalies. Anomaly detection approaches can be divided into two categories, i) searching a dictionary of known normal patterns and calculating distances (supervised learning methods), ii) deriving a normal pattern based on characteristics of the given samples (unsupervised learning methods). The problem of distinguishing of normal data points or sequences from anomalies is particularly difficult in complex domains such as stock market where time series do not follow a linear stochastic process. Previously, we developed a set of prediction models using some of the prominent existing supervised learning methods for fraud detection in securities market on a real dataset that is labeled based on litigation cases [42]. In that work, we adapted supervised learning algorithms to identify outliers (i.e. market manipulation samples) in stock market. We used a case study of manipulated stocks during 2003 that David Diaz introduced in his paper on analysis of stock market manipulation [43]. The dataset is manually labeled using SEC cases. Empirical results showed that Naïve Bayes outperformed other learning methods achieving an F_2 measure of 53% while the baseline F_2 measure was 17%. We extended the existing work on fraud detection in securities by adopting other algorithms, improving the performance results, identifying features that are misleading in the data mining process, and highlighting issues and weaknesses of these methods. The results indicate that adopting supervised learning algorithms for fraud detection in securities market using a labeled dataset is promising. However, there are two fundamental issues with the approach: first, it may be misleading to generalize such models to the entire domain as they are trained using one dataset, and second, using labeled datasets is impractical in the real world for many domains, especially securities market. This is because theoretically there are two approaches for evaluating outlier detection methods: i) using a labeled dataset, and ii) generating a synthetic dataset for evaluation. The standard approach in producing a labeled dataset for fraud detection in securities is using litigation cases to label observations as anomaly for a specific time and taking the rest of observations as normal. Accessing labeled datasets is a fundamental challenge in fraud detection and is impractical due to different costs associated to manually labeling data. It is a laborious and time consuming task, yet all existing literature on fraud detection in securities market using data mining methods, are based on this unrealistic approach [43] [44] [45] [46].

In an attempt to address challenges in developing an effective outlier detection method for non-parametric time series that are applicable to fraud detection in securities, we propose a prediction-based Contextual Anomaly Detection (CAD) method. Our method is different with the conventional prediction-based anomaly detection methods for time series in two aspects: i) the method does not require the assumption of time series being generated from a deterministic model (in fact as we indicated before, stock market time series are non-parametric and researchers have not been able to model these time series with reasonable accuracies to date [47]), and ii) instead of using a history of a given time series to predict its next consecutive values, we exploit the behavior of similar time series to predict the expected values.

The input to CAD is the set of time series $\{X_i | i \in \{1, 2, \dots, d\}\}$ from one sector such as S&P energy stocks (time series that are expected to have similar behavior as they share similar characteristics including underlying factors which determine the time series values) and a window size. First, a subset of time series is selected based on the window size, Second, a centroid is calculated representing the expected behavior of time series of the group within the window. The centroid is used along with statistical features of each time series X_i (e.g. correlation of the time series with the centroid) to predict the value of the time series at time t (i.e. $\hat{x}_{it}$). Table 2 describes the algorithm. This is a lazy approach, which uses the centroid along with other features of the time series for predicting the values of X_i:

$$\widehat{X_{it}} = \Psi(\Phi(X_t), c_t) + \varepsilon \quad (1)$$

where $\widehat{X_{it}}$ is the predicted values for the time series X_i at time t, $\Phi(X_t)$ is a function of time series features (e.g. the value of X_i at time stamp t-1, drift, auto regressive factor etc.), Ψ specifies the relationship of a given time series feature with the value of centroid at time t (i.e. c_t), and ε is the prediction error (i.e. $\sqrt{(\widehat{X_{it}} - X_{it})^2}$). The centroid time series C is the expected pattern (i.e. $E(X_1, X_2, \dots, X_d)$) which can be calculated by taking the mean or weighted mean of values of time series X_i at each time stamp t. We define Ψ as the inner product of statistical features of each time series and its correlation with the centroid. We use the *Pearson* correlation of each time series with the centroid to predict values of the time series because if the centroid correctly represents the pattern of time series in a group (i.e. industry sector), the correlation of individual time series with the centroid is an indicator of time series values. Third, we assign an anomaly score by taking the *Euclidean distance* of the predicted value and the actual value of the given time series (the threshold is defined by the standard deviation of each time series in the window). It has been shown that the Euclidean distance, although simple, outperforms many complicated distance measures and is competitive in the pool of distance measures for time series [35][48]. Moreover, the linear time complexity of Euclidean

Table 2. The Contextual Anomaly Detection Algorithm

Input: Time series $\{X_i | i \in \{1, 2, \dots, d\}\}$ from one sector, window size and overlap size (overlap is set to 4 data points in our experiments)

Return: Set of anomalies on each time series

```
1   start = overlap
2   while start < end of time – window size
3       start -= overlap
4       calculate the time series centroid C of {X_i | i ∈ {1, 2, ..., d}}
5       foreach i in {1, …, d}
6           c_i = correlation(X_i and C)
7           foreach data point x_i in X_i
8               predict point x̂_i based on c
9               if Euclidean distance(x_i, x̂_i) > std(X_i) then output x_i
10      start += window size
11  end while
```

distance makes it an ideal choice for large time series. Finally, we move the window and follow the same process. Figure 4 depicts the centroid time series within three time series of S&P energy sector with weekly frequency and a window size of 15 data points.

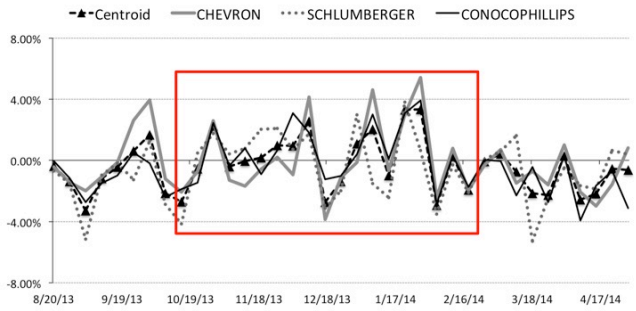

Figure 4. An example of three time series from S&P energy sector along their calculated centroid and a window size 15

A. Time Complexity

The problem of outlier detection in securities involves many time series with huge length. This makes the computational complexity of outlier detection methods important especially in the presence of High Frequency Trading[2] (HFT) where thousands of transactions are recorded per second in each time series (i.e. stock). The proposed method is linear with respect to the length of input time series. The centroid can be calculated in O(n) and using the Euclidean distance adds another O(n) to the computation leaving the overall computational complexity of the method in linear order (including other statistical features of a given time series such as drift and autoregressive factor in the predictive model will have the same effect on the computational complexity). However, there are constants such as the number of time series d and the number of local periods (e.g. 1-year periods that are used to capture outliers within that period of the original time series) that are multiplied to the total length of time series n. The constants are expected to be much smaller than the input size thus should not effect the order of computational complexity.

[2] HFT are algorithms that could submit many orders in millisecond. HFT accounts for 35% of the stock market trades in Canada and 70% of the stock trades in USA according to the 2010 Report on regulation of trading in financial instruments: Dark Pools & HFT.

B. Unlabeled Data and Injection of Outliers

We propose a systematic approach to synthesize data by injecting outliers in real securities market data that is known to be manipulation-free. The market data that we use - S&P constituents' data – is fraud-free (i.e. no market manipulation) thus considered outlier-free in the context of our problem. This is due to many reasons, most importantly, these stocks are:

- the largest companies in USA (with respect to their size of capital) and very unlikely to be *cornered*[3] by one party or a small group in the market,
- highly liquid (i.e. there are buyers and sellers at all times for the security and the buy/sell price-spread is small) thus practically impossible for a party to take control of a stock or affect the price in an arbitrary way,
- highly monitored and regulated both by the analysts in the market and regulatory organizations.

These are the major reasons which make S&P stocks a reliable benchmark for risk analysis, financial forecasting and fraud detection with a long history in industry and in numerous research works [49] [50] [46].

In our proposed approach, values of synthetic outliers for a given time series are generated based on the distribution of subsequences of the given time series (e.g. in periods of 1 year). It is important to note that our proposed outlier detection method follows a completely different mechanism and is not affected by the process of outlier injection in any way (we elaborate more on this at the end of this section). The conventional approach in defining outliers for a normal distribution $N(\mu, \sigma^2)$, is taking observations with distance of three standard deviation from the mean (i. e. $\mu \pm 3\sigma$) as outliers. However, when the distribution is skewed we need to use a different model to generate outliers. We adopted Tukey's method [51] for subsequences that do not follow a normal distribution. It has been shown that Tukey's definition for outliers is an effective approach for skewed data [52]. Formally, we propose generating artificial outliers using the following two-fold model:

$$\tau(x_{it}) = \begin{cases} \mu + [Q_3 + (3 * IQR)] \; or \; \mu - [Q_1 - (3 * IQR)] & if \; \gamma_1 > \varepsilon \\ \mu \pm 3\sigma & if \; N(\mu, \sigma^2) \end{cases} \quad (2)$$

where Q_1 is the lower quartile (25th percentile), Q_3 is the upper quartile (75th percentile), IQR represents the inter-quartile (i.e. Q_3 - Q_1) of the data, and γ_1 represents the skewness or third moment of the data distribution:

[3] Cornering a security means taking control of the majority of the asset in a way that the owner can affect the price by control over supply.

$$\gamma_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\sum_1^k (x_i-\mu)^3}{n} \quad (3)$$

and k is the length of the subsequence of time series X_i (i.e. number of data points in the subsequence). γ_1 is 0 for a normal distribution as it is symmetric. The values in a given time series are randomly substituted with the synthetic outliers $\tau(x_{it})$.

We emphasize that the process of injecting outliers to create synthesized data using the real market data is completely separate from our anomaly detection process. Anomalies are injected randomly and this information is not used in the proposed anomaly detection process. The injected outliers in a time series are based solely on the time series itself and not the group of time series. Furthermore, the outlier detection method that we propose is an unsupervised learning method and the ground truth that is based on the synthetic data, is only used to evaluate performance of the proposed method and the competitive methods after capturing outliers. Injecting anomalies for evaluating outlier detection methods has been attempted in different domains such as intrusion detection [53]. One may ask, assuming the above model defines outliers, can we use this same two-fold model approach to identify outliers for a given set of time series? The answer is no, because the statistical characteristics of the time series such as mean, standard deviation and skewness are affected by outliers, therefore, these values may be misleading as the input time series include outliers.

We use the market data from S&P constituents datasets that are considered outlier-free. The process to synthesize artificial outliers described in Section 3.2 is used to inject outliers in the real datasets. These datasets are used as the input data for the outlier detection methods in our experiments. We use the performance measures *precision*, *recall* and *F-measure* in our experiments. If the null hypothesis is that all and only the outliers are retrieved, absence of type I and type II errors correspond to maximum *precision* (no false positives) and maximum recall (no false negatives) respectively. *Precision* is a measure of exactness or quality, whereas *recall* is a measure of completeness or quantity.

We compare performance of the proposed method with two competing algorithms for time series anomaly detection, Naïve predictor (Random walk) and kNN. In this paper we identified three criteria for effective anomaly detection methods in stock market: i) have O(n) or close to linear time complexity, ii) be able to detect individual anomalous data points, iii) rely on an unsupervised learning approach. The proposed method is designed to satisfy these criteria. Random walk and kNN are carefully selected as competing methods satisfying these criteria. Random walk is a widely accepted benchmark for evaluating time-series forecasting [54], which predicts x_{t+1} through a random walk (a jump) from x_t. Random walk is equivalent to ARIMA (0,1,0) (Auto-Regressive Integrated Moving Average) [55]. This model does not require the stationary assumption for time series, however, assumes that the time series follow a first-order Markov process (because the value of X_{t+1} depends only on the value of X at time t). x_{t+1} is anomalous if it is significantly deviated from its prediction. We use kNN as a proximity based approach for outlier detection. Furthermore, kNN, although simple, reached promising results in the work on detecting stock market manipulation in a pool of different algorithms including decision trees, Naïve Bayes, Neural Networks and SVM. For each data point p we calculate $D^k(p)$ as the distance of all other k[th] nearest points (using Euclidean distance). A data point p would be anomalous if $D^k(p)$ is significantly different from other data pints q with $D^k(q)$.

IV. DATA

We use several datasets from different industry sectors of S&P 500 constituents (see Appendix A for more information on S&P sectors). We use these datasets in two different granularity of daily and weekly frequencies. The S&P 500 index includes the largest market cap stocks that are selected by a team of analysts and economists at Standard and Poor's[4]. The S&P 500 index is the leading indicator of US equities and reflects the characteristics of top 500 largest market caps. As we indicated in Section 3.2 these stocks (time series) are assumed to have no anomalies (i.e. no manipulations), as they are highly liquid and closely monitored by regulatory organizations and market analysts. We use 10 different datasets including 636 time series over a period of 40 years. To the best of our knowledge, this study surpasses the previous works in terms of both the duration and the number of time series in the datasets. Table 3 describes a list of datasets that we extracted from Thompson Reuters database for experiments to study and validate our proposed method (the CSV files are available at www.ualberta.ca/~golmoham/DSAA2015/). The table includes the total number of data points with a finite value (excluding NaN) in each dataset. These time series are normalized (by taking the percentage change) in a preprocessing step of our data mining process. Normalizing and scaling features before the outlier detection process is crucial. This is also a requirement for many statistical and machine learning methods. For example, consider the price, which is the most important feature that should be monitored for detecting market manipulation in a given security. The price of a security would include the trace of market manipulation activities because any market manipulation scheme seeks profit from deliberate change in price of that security. However, the price of a stock does not reflect the size of a company nor the revenue. Also, the wide range of prices is problematic when taking the first difference of the prices. A standard approach is using the price percentage change (i.e. return), $R_t = (P_t - P_{t-1})/P_{t-1}$ where R_t and P_t represent return

Table 3. List of datasets for experiments on outlier detection methods for fraud detection in securities

S&P Sector	Number of time series	Number of data points [weekly frequency]	Number of data points [daily frequency]
Energy	44	63,000 +	315,000 +
Financials	83	117,000 +	587,000 +
Consumer Discretionary	85	111,000 +	558,000 +
Information Technology	66	80,000 +	395,000 +
Consumer Staples	40	64,000 +	323,000 +

[4] Standard and Poor is an American financial services and credit rating agency that has been publishing financial research and analysis on stocks and bonds for over 150 years.

and price of the security at time t respectively. The sample space of R_t is $[-1, M]$ and $M > 0$. The ratio of artificial outliers that are injected in the outlier-free dataset (see section 3.2) is 0.001 of the total number of data points in each dataset.

V. RESULTS

The conventional performance measures are not appropriate for anomaly detection because the misclassification costs are unequal. The second issue which makes performance evaluation challenging is unbalanced classes. Anomaly detection for detecting stock market manipulation encompasses both properties because i) false negatives are more costly as missing a market manipulation period by predicting it to be normal, hurts performance of the method more than including a normal case by predicting it to be market manipulation, ii) the number of market manipulations (i.e. anomalies) constitute a tiny percentage of the total number of transactions in the market. We argue the performance evaluation, thus performance measures should be only based on predicting anomalies and avoid including results of predicting normal data points in the performance evaluation. Therefore, we only report numbers on predicting anomalies[5]. We use F-measures with higher β values to give higher weights to recall of correctly identifying anomalies:

$$F_\beta = (1+\beta^2) * \frac{P * R}{(\beta^2 * P) + R} = \frac{(1+\beta^2) * TP}{(1+\beta^2) * TP + (\beta^2 * FP) + FP} \quad (4)$$

where P and R represent the *precision* and *recall* respectively ($P = \frac{TP}{TP+FP}$ and $R = \frac{TP}{TP+FN}$), TP is true positive (the number of outliers predicted correctly as outliers), FP is false positive (the number of normal data points that are predicted as outliers), TN is true negative (the number of normal data points that are predicted as normal), FN is false negative (the number of outliers that are incorrectly predicted as normal), and $\beta \in \mathbb{N}$ and $\beta > 0$.

We run experiments with different window sizes (15, 20, 24, 30 and 35) on all 10 datasets (5 industry sectors with daily and weekly frequencies. See Table 3.). Figure 5 illustrates the average performance results over all window sizes of each anomaly detection method on each dataset with daily frequency. As can be noted in the table in the Appendix the results are stable regardless of the window size in the experiments. Our approach, CAD, clearly outperforms the other two methods on recall (i.e. it is superior at finding anomalies). A hypothetical predictor that predicts all data points as anomalies would reach an F_4-measure of 0.016 since the injected outliers only represent 0.001 of the total number of data points. Our objective is maximizing recall without compromising precision. The precision is less than 0.5% for all three algorithms while CAD reaches much higher recall in predicting anomalies. The baseline for precision (by predicting all data points as anomalies) is less than 0.04% because the total number of anomalies constitutes less than 0.1% of data, which drops to 0.04% after data preprocessing. Although, avoiding false positives is generally desirable, it is not the focus in detecting

[5] The complete report of results including performance evaluation for both normal and anomaly periods see https://gist.github.com/koosha/a2457ce63feb41ef3ea4

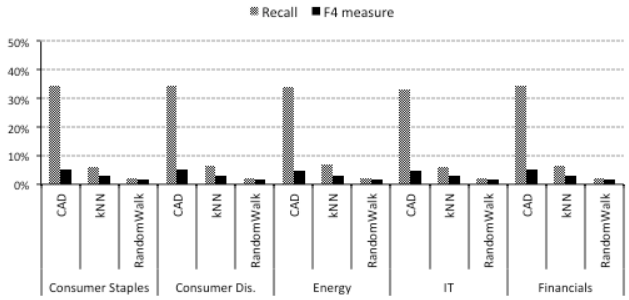

Figure 5. Average recall and F_4-measure on daily data of S&P sectors

stock market manipulation because missing an anomaly (potential market manipulation) hurts the method much more than incorrectly predicting a sample as anomalous (false positive). We emphasize that the objective of this paper is improving recall without compromising the precision measures using other applicable methods (precision of kNN and Random Walk is less than 0.5%). CAD improves recall from 7% to 33% without compromising the precision.

Figure 6 describes the average recall and F_4-measure of the anomaly detection methods on each dataset with weekly frequency. It shows a similar trend where CAD clearly outperforms its contenders.

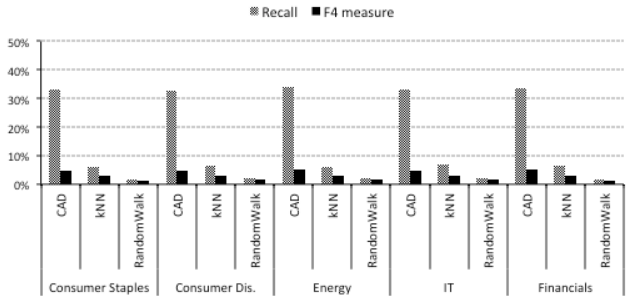

Figure 6. Average recall and F_4-measure on weekly data of S&P sectors

VI. CONCLUSION

We proposed a Contextual Anomaly Detection (CAD) method for complex time series that is applicable for identifying stock market manipulation. The method considers not only the context of a time series in a time window but also the context of similar time series in a group of similar time series. We designed and implemented a comprehensive set of experiments to evaluate the proposed method on 5 different S&P sectors with daily and weekly frequencies over the past 40 years. The results indicate that the proposed method outperforms kNN and Random Walk in identifying anomalies in time series grouped by sectors. Although many anomalies have been established (relatively high recall), our method still flags false positives (low precision). This means that regulators would have to sift through the true and false positives. As a future work, a second phase would consist of weeding out some of the false positives by means of a classifier to improve the precision. The same idea has been successfully applied to eliminate falsely detected nodules by

a computer vision technique in pulmonary scans [3]. The problem we addressed is a challenging problem because we attempt to detect anomalies in an unsupervised way in times series where no deterministic function can model data.

ACKNOWLEDGMENTS

We acknowledge the financial support from Natural Science and Engineering Research Council (NSERC), Alberta Innovates Centre for Machine Learning (AICML) and Alberta Innovates Technology Futures (AITF). We acknowledge using *Pandas*[6], *Numpy*[7] and *scikit-learn*[8] for implementing the experiments and using the high performance computing platform of WestGrid[9].

REFERENCES

[1] A. Fabrizio, R. B. Zohary, and L. Feo, "Outlier Detection Using Default Logic," in *the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI)*, 2003, pp. 833–838.

[2] P. D. L. Ertöz, E. Eilertson, A. Lazarevic, P.-N. Tan, V. Kumar, J. Srivastava, *Minds-minnesota intrusion detection system*, Next Gener. MIT Press, 2004.

[3] C. Spence, L. Parra, and P. Sajda, "Detection, Synthesis and Compression in Mammographic Image Analysis with a Hierarchical Image Probability Model," p. 3, Dec. 2001.

[4] E. Aleskerov, B. Freisleben, and B. Rao, "CARDWATCH: a neural network based database mining system for credit card fraud detection," in *Proceedings of the IEEE/IAFE 1997 Computational Intelligence for Financial Engineering (CIFEr)*, 1997, pp. 220–226.

[5] V. Chandola, a. Banerjee, and V. Kumar, "Anomaly Detection for Discrete Sequences: A Survey," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 823–839, May 2012.

[6] P. D. Ashok N. Srivastava, "Discovering System Health Anomalies Using Data Mining Techniques," in *the Joint Army Navy NASA Air Force Conference on Propulsion*, 2005.

[7] J. Lin, E. Keogh, A. Fu, and H. Herle, "Approximations to Magic: Finding Unusual Medical Time Series," in *18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)*, 2005, pp. 329–334.

[8] X. X. Li Wei, Eamonn Keogh, "SAXually Explicit Images: Finding Unusual Shapes," in *the 2006 IEEE international conference on data mining*, 2006, pp. 18–22.

[9] D. Yankov, E. Keogh, and U. Rebbapragada, "Disk aware discord discovery: finding unusual time series in terabyte sized datasets," *Knowl. Inf. Syst.*, vol. 17, no. 2, pp. 241–262, Mar. 2008.

[10] Z. Ferdousi and A. Maeda, "Unsupervised Outlier Detection in Time Series Data," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, 2006, pp. x121–x121.

[11] E. Keogh, J. Lin, and A. Fu, "HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence," in *Fifth IEEE International Conference on Data Mining (ICDM'05)*, 2005, pp. 226–233.

[12] E. Keogh, J. Lin, S. Lee, and H. Van Herle, "Finding the most unusual time series subsequence: algorithms and applications," *Knowl. Inf. Syst.*, vol. 11, no. 1, pp. 1–27, 2007.

[13] M. Minenna, "The Detection of Market Abuse on Financial Markets: A Quantitative Approach," *Quad. di Finanz.*, vol. 54, 2003.

[14] Y. Song, L. Cao, X. Wu, G. Wei, W. Ye, and W. Ding, "Coupled behavior analysis for capturing coupling relationships in group-based market manipulations," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, 2012, p. 976.

[15] Golmohammadi, K. and Zaiane, O.R. 2012. Data Mining Applications for Fraud Detection in Securities Market. 2012 European Intelligence and Security Informatics Conference (Aug. 2012), 107–114.

[16] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing SAX: a novel symbolic representation of time series," *Data Min. Knowl. Discov.*, vol. 15, no. 2, pp. 107–144, Apr. 2007.

[17] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, "Fast subsequence matching in time-series databases," *ACM SIGMOD Rec.*, vol. 23, no. 2, pp. 419–429, Jun. 1994.

[18] J. Zhang, F.-C. Tsui, M. M. Wagner, and W. R. Hogan, "Detection of outbreaks from time series data using wavelet transform.," in *AMIA Annual Symposium Proceedings*, 2003, pp. 748–752.

[19] V. Chandola, D. Cheboli, and V. Kumar, "Detecting anomalies in a time series database," in *Technical Report*, 2009, p. 12.

[20] J. Ma and S. Perkins, "Time-series novelty detection using one-class support vector machines," in *Proceedings of the International Joint Conference on Neural Networks, 2003.*, 2003, vol. 3, pp. 1741–1745.

[21] U. Rebbapragada, P. Protopapas, C. E. Brodley, and C. Alcock, "Finding anomalous periodic time series," *Mach. Learn.*, vol. 74, no. 3, pp. 281–313, Dec. 2008.

[22] P. Protopapas, J. M. Giammarco, L. Faccioli, M. F. Struble, R. Dave, and C. Alcock, "Finding outlier light curves in catalogues of periodic variable stars," *Mon. Not. R. Astron. Soc.*, vol. 369, no. 2, pp. 677–696, Jun. 2006.

[23] C. Chatfield, *The Analysis of Time Series: An Introduction*. Chapman and Hall/CRC; 6 edition, 2003.

[24] F. Knorn and D. J. Leith, "Adaptive Kalman Filtering for anomaly detection in software appliances," in *IEEE INFOCOM 2008 - IEEE Conference on Computer Communications Workshops*, 2008, pp. 1–6.

[25] A. B. C. Junshui Ma, Simon Perkins, "Online Novelty Detection on Temporal Sequences," in *the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 613–618.

[26] C. C. Michael and A. Ghosh, "Two state-based approaches to program-based anomaly detection," in *Proceedings 16th Annual Computer Security Applications Conference (ACSAC'00)*, 2000, pp. 21–30.

[27] T. Lotze, G. Shmueli, S. Murphy, and H. Burkom, "A wavelet-based anomaly detector for early detection of disease outbreaks," *Work. Mach. Learn. Algorithms Surveill. Event Detect. 23rd Intl Conf. Mach. Learn.*, 2006.

[28] Z. Liu, J. X. Yu, and L. Chen, "Detection of Shape Anomalies: A Probabilistic Approach Using Hidden Markov Models," in *2008 IEEE 24th International Conference on Data Engineering*, 2008, pp. 1325–1327.

[29] Y. Qiao, X. W. Xin, Y. Bin, and S. Ge, "Anomaly intrusion detection method based on HMM," *Electronics Letters*, vol. 38, no. 13. IET Digital Library, pp. 663–664, 20-Jun-2002.

[30] X. Zhang, P. Fan, and Z. Zhu, "A new anomaly detection method based on hierarchical HMM," in *Proceedings of the 8th International Scientific and Practical Conference of Students, Post-graduates and Young Scientists. Modern Technique and Technologies. MTT'2002 (Cat. No.02EX550)*, 2003, pp. 249–252.

[31] P. K. Chan and M. V. Mahoney, "Modeling Multiple Time Series for Anomaly Detection," *Fifth IEEE Int. Conf. Data Min.*, pp. 90–97, 2005.

[32] M. V. M. and P. K. Chan., "Trajectory boundary modeling of time series for anomaly detection," in *KDD Workshop on Data Mining Methods for Anomaly Detection*, 2005.

[6] http://pandas.pydata.org/

[7] http://www.numpy.org/

[8] http://scikit-learn.org/stable/

[9] www.westgrid.ca

[33] S. Salvador and P. Chan, “Learning States and Rules for Detecting Anomalies in Time Series,” *Appl. Intell.*, vol. 23, no. 3, pp. 241–255, Dec. 2005.

[34] P. Protopapas, J. M. Giammarco, L. Faccioli, M. F. Struble, R. Dave, and C. Alcock, “Finding outlier light curves in catalogues of periodic variable stars,” *Mon. Not. R. Astron. Soc.*, vol. 369, no. 2, pp. 677–696, Jun. 2006.

[35] R. Giusti and G. E. A. P. A. Batista, “An Empirical Comparison of Dissimilarity Measures for Time Series Classification,” in *2013 Brazilian Conference on Intelligent Systems*, 2013, pp. 82–88.

[36] B. Pincombe, “Anomaly Detection in Time Series of Graphs using ARMA Processes,” vol. 24, no. 4, pp. 2–10, 2005.

[37] H. Zare Moayedi and M. A. Masnadi-Shirazi, “Arima model for network traffic prediction and anomaly detection,” in *2008 International Symposium on Information Technology*, 2008, vol. 4, pp. 1–6.

[38] D. Moore, *Introduction to the Practice of Statistics*. 2004.

[39] C. E. R. and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge: MIT Press., 2006.

[40] L. Rabiner and B. Juang, “An introduction to hidden Markov models,” *IEEE ASSP Mag.*, vol. 3, no. 1, pp. 4–16, 1986.

[41] V. Jecheva, “About Some Applications of Hidden Markov Model in Intrusion Detection Systems,” in *International Conference on Computer Systems and Technologies*, 2006.

[42] Golmohammadi, K., Zaiane, O.R. and Díaz, D. 2014. Detecting Stock Market Manipulation using Supervised Learning Algorithms. The 2014 International Conference on Data Science and Advanced Analytics (DSAA’2014).

[43] D. Diaz, B. Theodoulidis, and P. Sampaio, “Analysis of stock market manipulations using knowledge discovery techniques applied to intraday trade prices,” *Expert Syst. Appl.*, vol. 38, no. 10, pp. 12757–12771, Sep. 2011.

[44] J. D. Kirkland, T. E. Senator, J. J. Hayden, T. Dybala, H. G. Goldberg, and P. Shyr, “The NASD Regulation Advanced-Detection System (ADS),” *AI Magazine*, vol. 20, no. 1. p. 55, 15-Mar-1999.

[45] T. E. Senator, “Ongoing management and application of discovered knowledge in a large regulatory organization,” in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’00*, 2000, pp. 44–53.

[46] H. Öğüt, M. Mete Doğanay, and R. Aktaş, “Detecting stock-price manipulation in an emerging market: The case of Turkey,” *Expert Syst. Appl.*, vol. 36, no. 9, pp. 11944–11949, Nov. 2009.

[47] Y. Wang, “Mining stock price using fuzzy rough set system,” *Expert Syst. Appl.*, vol. 24, no. 1, pp. 13–23, Jan. 2003.

[48] E. Keogh and S. Kasetty, “On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration,” *Data Min. Knowl. Discov.*, vol. 7, no. 4, pp. 349–371, Oct. 2003.

[49] D. Enke and S. Thawornwong, “The use of data mining and neural networks for forecasting stock market returns,” *Expert Syst. Appl.*, vol. 29, no. 4, pp. 927–940, Nov. 2005.

[50] W. Huang, Y. Nakamori, and S.-Y. Wang, “Forecasting stock market movement direction with support vector machine,” *Comput. Oper. Res.*, vol. 32, no. 10, pp. 2513–2522, Oct. 2005.

[51] J. W. Tukey, *Exploratory Data Analysis*. 1977.

[52] S. Seo, “A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets,” in *Thesis*, 2006.

[53] W. Fan, M. Miller, S. Stolfo, W. Lee, and P. Chan, “Using artificial anomalies to detect unknown and known network intrusions,” *Knowl. Inf. Syst.*, vol. 6, no. 5, pp. 507–527, Apr. 2004.

[54] O. Gjolberg and B.-A. Bengtsson, “Forecasting quarterly hog prices: Simple autoregressive models vs. naive predictions,” *Agribusiness*, vol. 13, no. 6, pp. 673–679, Nov. 1997.

[55] G. C. R. George E. P. Box, Gwilym M. Jenkins, *Time Series Analysis: Forecasting and Control*, 4th ed. 2008.

APPENDIX

Below is the performance results of the proposed Contextual Anomaly Detection (CAD) method, kNN and Random Walk in predicting anomalies on all datasets with daily frequency. More experimental results, including results on these datasets with weekly frequency can be found at https ://gist.github.com/koosha/a2457ce63feb41ef3ea4.

win size	Dataset	Algorithm	Precision (%)	Recall (%)	F_2 measure (%)	F_4 measure (%)
15	Consumer Staples	CAD	0.33	34.70	1.59	4.86
		kNN	0.28	6.02	1.17	2.71
		RandomWalk	0.24	1.65	0.75	1.22
	Consumer Dis.	CAD	0.33	34.15	1.60	4.88
		kNN	0.29	6.26	1.24	2.86
		RandomWalk	0.25	1.72	0.79	1.28
	Energy	CAD	0.33	34.49	1.58	4.83
		kNN	0.29	6.36	1.23	2.86
		RandomWalk	0.34	2.39	1.09	1.77
	IT	CAD	0.34	33.69	1.63	4.98
		kNN	0.33	6.83	1.40	3.19
		RandomWalk	0.32	2.14	1.00	1.60
	Financials	CAD	0.34	35.47	1.65	5.05
		kNN	0.34	7.18	1.42	3.27
		RandomWalk	0.38	2.62	1.20	1.94
20	Consumer Staples	CAD	0.33	34.02	1.60	4.88
		kNN	0.25	5.42	1.07	2.46
		RandomWalk	0.31	2.12	0.98	1.58
	Consumer Dis.	CAD	0.32	34.16	1.53	4.69
		kNN	0.31	6.94	1.31	3.06
		RandomWalk	0.35	2.53	1.12	1.85
	Energy	CAD	0.31	32.30	1.48	4.53
		kNN	0.31	6.77	1.30	3.03
		RandomWalk	0.33	2.29	1.04	1.69
	IT	CAD	0.34	34.01	1.63	4.97
		kNN	0.32	6.67	1.34	3.07
		RandomWalk	0.37	2.52	1.17	1.88
	Financials	CAD	0.34	34.62	1.62	4.95
		kNN	0.31	6.61	1.30	3.00
		RandomWalk	0.28	1.96	0.89	1.45
24	Consumer Staples	CAD	0.36	35.01	1.72	5.23
		kNN	0.31	6.42	1.31	2.99
		RandomWalk	0.30	1.93	0.92	1.45
	Consumer Dis.	CAD	0.36	34.19	1.71	5.18
		kNN	0.34	6.84	1.41	3.21
		RandomWalk	0.40	2.63	1.25	1.99
	Energy	CAD	0.36	35.77	1.72	5.23
		kNN	0.35	7.35	1.47	3.37
		RandomWalk	0.33	2.19	1.03	1.65
	IT	CAD	0.34	33.96	1.63	4.97
		kNN	0.23	4.94	0.99	2.27
		RandomWalk	0.35	2.36	1.10	1.76
	Financials	CAD	0.33	34.46	1.61	4.93
		kNN	0.31	6.70	1.31	3.03
		RandomWalk	0.33	2.25	1.04	1.68
30	Consumer Staples	CAD	0.35	34.21	1.68	5.11
		kNN	0.33	6.80	1.38	3.15
		RandomWalk	0.34	2.27	1.06	1.70
	Consumer Dis.	CAD	0.35	33.54	1.70	5.15
		kNN	0.32	6.43	1.34	3.03
		RandomWalk	0.33	2.09	1.01	1.59
	Energy	CAD	0.35	34.57	1.70	5.18
		kNN	0.35	7.20	1.47	3.35
		RandomWalk	0.29	1.94	0.91	1.45
	IT	CAD	0.32	31.83	1.54	4.68
		kNN	0.28	5.98	1.20	2.75
		RandomWalk	0.34	2.31	1.07	1.72
	Financials	CAD	0.36	34.13	1.74	5.27
		kNN	0.32	6.38	1.34	3.02
		RandomWalk	0.34	2.18	1.05	1.66
35	Consumer Staples	CAD	0.36	34.13	1.71	5.19
		kNN	0.29	5.86	1.20	2.73
		RandomWalk	0.33	2.18	1.04	1.65
	Consumer Dis.	CAD	0.39	35.29	1.85	5.60
		kNN	0.35	6.85	1.46	3.28
		RandomWalk	0.37	2.28	1.12	1.75
	Energy	CAD	0.32	32.55	1.55	4.73
		kNN	0.32	6.97	1.36	3.16
		RandomWalk	0.27	1.85	0.85	1.38
	IT	CAD	0.35	33.07	1.67	5.07
		kNN	0.31	6.23	1.29	2.92
		RandomWalk	0.34	2.22	1.06	1.68
	Financials	CAD	0.36	32.68	1.73	5.21
		kNN	0.29	5.65	1.21	2.72
		RandomWalk	0.37	2.32	1.13	1.77