**Review of Methods of Anomaly Detection in Time Series Data**

Existing anomaly detection methods for stock market data can be classified based on how they transform the data prior to anomaly detection and their process of identifying anomalies. (Summary below adapted from Golmohammadi 2015)

**Transformations:** Before anomaly detection begins, data must be transformed in order to handle high dimensionality, scaling, and noise and to achieve computational efficiency. This can be done in three ways: aggregation, discretization, and signal processing.

***Aggregation*** involves dimensionality reduction by aggregating consecutive values, typically by representing them by their average.

***Discretization*** involves converting the time series into a discrete sequence of finite alphabets, which allows you to use existing symbolic sequence anomaly detection algorithms and also to improve computational efficiency.

***Signal processing*** involves mapping the data to a different space in order to make it easier to detect outliers and to reduce dimensionality.

**Processes of Identifying Anomalies:** The current processes of identifying anomalies can be categorized into five groups: window based, proximity based, prediction based, hidden markov model (HMM) based, and segmentation based.

***Window Based:*** The time series is divided into evenly sized windows of subsequences and the distance from that sliding window to the windows in the training database determines the anomaly score. Selection of the optimal window size must be done carefully and take into account the length of the anomalous subsequence. This method can be computationally expensive (*O((nl)2)*), where *n* is the number of samples in the training and testing datasets and *l* is the average length of the time series.

***Proximity Based:*** This method uses pairwise proximity between the testing and training time series using an appropriate distance/similarity kernel. The similarity measure is then used to measure the distance of every two given sequences. A k-NN or clustering method is used to calculate the anomaly score. A major disadvantage is that this method can identify an anomalous time series, but it cannot pinpoint the exact location of the anomaly. It is also highly sensitive to the similarity measure used.

***Prediction Based:*** These assume that the normal time series is generated from a statistical process while the anomalies do not fit this process. However, the length of history used for prediction is very influential in locating anomalies. In addition, these methods perform very poorly when the time series was not generated from a statistical process.

***Hidden Markov Model (HMM) Based:*** In these models, the training dataset is used to build a hidden Markov model (HMM), which is then used to probabilistically assign an anomaly score to a given test time series. However, if the underlying time series is not generated from an HMM, it will perform poorly.

***Segmentation Based:*** The time series is divided into segments. These methods assume that there is an underlying Finite State Automaton (FSA) that models the normal time series, and anomalies are detected when segments do not fit the FSA. However, the segmentation procedure may obscure anomalies.

**Table 1. Anomaly Detection Methods for Time Series Data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Transformation** | | |
|  |  | **Aggregation** | **Discretization** | **Signal Processing** |
| **Technique** | **Window Based** | kNN (Chandola 2009)  SVM (Ma 2003b)  (Golmohammadi 2015) | kNN (Chandola 2009) |  |
| **Proximity Based** | PCAD (Protopapas 2006, Rebbapragada 2009)  Martingale (Fedorova 2012, Ho 2010, Vovk 2003) |  |  |
| **Prediction Based** | Moving Average (Chatfield 2004)  Auto Regression (Chatfield 2004)  Kalman Filters (Knorn 2008)  SVM (Ma 2003a) | FSA (Michael 2000) | Wavelet (Lotze 2006, Zhang 2003) |
| **HMM Based** | (Liu 2008) | (Qiao 2002)  (Zhang 2003) |  |
| **Segmentation** | (Chan 2005)  (Mahoney 2005)  (Salvador 2005) |  |  |

**References**

Chan, P., & Mahoney, M. (2005). Modeling multiple time series for anomaly detection. *Data Mining, Fifth IEEE International Conference on,* 8 pp.

Chandola, V., Cheboli, D., & Kumar, V. (2009) Detecting anomalies in a time series database. *Technical Report*, 1-12.

Chatfield, C. (2004). *The analysis of time series : An introduction* (6th ed., Texts in statistical science). Boca Raton, FL: Chapman & Hall/CRC.

Fedorova, V., Gammerman, A., Nouretdinov, I., & Vovk, V. (2012). Plug-in martingales for testing exchangeability on-line. *Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK.*

Golmohammadi, K., & Zaiane, O. (2015). Time series contextual anomaly detection for detecting market manipulation in stock market. *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on,* 1-10.

Ho, S-S., & Wechsler, H. (2010). A Martingale Framework for Detecting Changes in Data Streams by Testing Exchangeability. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* *32*(12), 2113-2127.

Knorn, F., & Leith, D. (2008). Adaptive Kalman Filtering for anomaly detection in software appliances. *INFOCOM Workshops 2008, IEEE,* 1-6.

Liu, Z., Yu, J. X., Chen, L., & Wu, D. (2008). Detection of shape anomalies: A probabilistic approach using hidden Markov Models. 1325-1327.

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

Ma, Junshui, & Perkins, Simon. (2003a). Online novelty detection on temporal sequences. 613-618.

Ma, J., & Perkins, S. (2003b). Time-series novelty detection using one-class support vector machines. *Neural Networks, 2003. Proceedings of the International Joint Conference on,* *3*, 1741-1745.

Mahoney, M., & Chan, P. (2005) Trajectory boundary modeling of time series for anomaly detection. *KDD Workshop on Data Mining Methods for Anomaly Detection*.

Michael, C., & Ghosh, A. (2000) Two state-based approaches to program-based anomaly detection. *Proceedings 16th Annual Computer Security Applications Conference (ACSAC’00)*, 21-30.

Protopapas, P., Giammarco, J., Faccioli, L., Struble, M., Dave, R., & Alcock, C. (2006). Finding outlier light curves in catalogues of periodic variable stars. *Monthly Notices of the Royal Astronomical Society,* *369*(2), 677-696.

Rebbapragada, U., Protopapas, P., Brodley, C., & Alcock, C. (2009) Finding anomalous periodic time series. *Machine Learning*, 74(3), 281-313.

Salvador, S., & Chan, P. (2005). Learning states and rules for detecting anomalies in time series. *Applied Intelligence,* *23*(3), 241-255.

Qiao, Xin, Bin, & Qiao, Y. (2002). Anomaly intrusion detection method based on HMM. *Electronics Letters,* *38*(13), 663-664.

Vovk, V., Nouretdinov, I., Gammerman, A. (2003) Testing exchangeability on-line. *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC.*

Zhang, X., Fan, P., & Zhu, Z. (2003). A new anomaly detection method based on hierarchical HMM. *Parallel and Distributed Computing, Applications and Technologies, 2003. PDCAT'2003. Proceedings of the Fourth International Conference on,* 249-252.

Zhang, J., Tsui, F., Wagner, M., & Hogan, W. (2003). Detection of outbreaks from time series data using wavelet transform. *AMIA Annual Symposium Proceedings. AMIA Symposium,* 748-52.