Abnormality Detection in Stock Market with Unsupervised Learning

Tom Han, Linguo Ren, Professor David Byrd
CSCI 3456 Financial Machine Learning
Computer Science Department, Bowdoin College, Brunswick, Maine

Introduction

In the context of the stock market, **abnormalities** refer to outliers or unusual patterns that deviate from the expected behavior of stock prices. These abnormalities can occur due to various factors, including sudden market news, economic events, or irregular trading activities. Detecting these outliers is crucial for investors and analysts to identify potential risks or opportunities. However, the task is challenging due to the high volume of data, rapid market changes, and the subtle nature of these abnormalities.

Detecting stock market abnormalities is a crucial yet challenging task in the financial sector. Traditional statistical methods, technical indicators (moving average, RSI, etc.), and directly clustering have underperformed in stock data [2]. The complexity arises from serval factors:

- **High volume of data**: Around 1 million transactions took place in the most traded stock tickers per day.
- **Rapid and repetitive market changes:** High frequency trading (HFT) and market microstructure effects causes stock data at the time horizon of seconds to microseconds to oscillate.
- **Subtle nature of abnormality:** With the high volumes of transaction, stock price usually changes around ~$2 per day.
- **Time series data**: Stock prices are a timeseries where neighboring observations are highly correlated. For factor analysis and clustering algorithms, which requires each variable to be uncorrelated, this results in poor performance.

In this project, we explored the possibility of using a **temporal convolutional based autoencoder (TCN-AE)** to detection abnormality in stock prediction [1]. An autoencoder performs dimension reduction which removes the correlation between neighboring observations. Compares to traditional autoencoder, TCN-AE have significant better performance in high frequency periodic data like ECG data [4]. We reconstructed the TCN-AE structure in PyTorch and trained out model on resampled (100ms per observation) market data and ran clustering algorithm for abnormality detection.

Project Goals & Questions

- Using unsupervised learning to detect abnormalities in stock data.
- Implementing a Temporal-Convolution Network Autoencoder (TCN-AE) to reduce size and dimension of unlabeled high frequency data efficiently.
- Testing different hyperparameters such as the distance metric and linkage criteria for clustering.
- Identifying repeating and overlapping abnormal data points by experimenting with different combinations of distance metrics and linkage criteria.

Future Directions

- **More Training Dataset:** Since the dataset from lobster data is limited to a week, our model may not yet be capable of capturing insider trading effectively. However, our model was still able to identify abnormal data segments of market hours. With a larger dataset and training hours, the model could be used to detect abnormalities like insider trading more accurately. A well-trained model could not only help organizations like the SEC detect insider trading, but also serve as an effective tool for individual investors to assess the stock market.
- **Finer abnormality detection with reconstruction error:** In Thill et al, 2019, an algorithm for precise (millisecond level) abnormality segment was introduced. Which can be a subsequent step after identifying the abnormal interval (30min).

Methods

Building a Temporal Convolutional Autoencoder (TCN-AE)

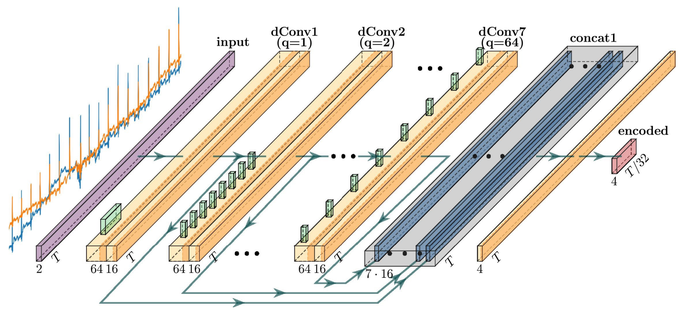
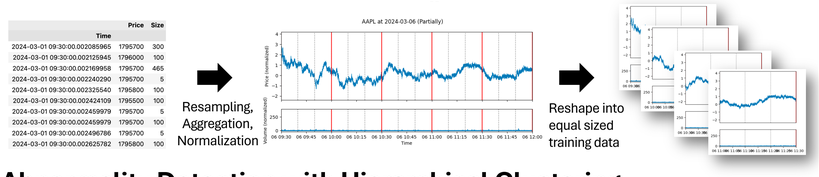


Figure 1. Encoder Structure of TCN-AE. Figure cited from Thill et al, 2021. 7 layers of "temporal blocks" with one dilated convolution layer of 64 channels followed by 1x1 convolution to reduce channel size to 16. Outputs of each temporal block is passed into the subsequent temporal block, A skip layer concatenating all outputs is passed into a final convolution and average pooling later to encodes the data into a 4-channel encoded form. In the decoder structure, an up-sampling pool followed by 7 layers of temporal blocks of similar structure is used to reconstruct the input data. (Implementation not show on this figure). The final encoded data have 4 channels of 360 observations (from 2 x 18000).

Data Preprocessing

We collected high frequency trading data gathered from lobster data [7] and resampled it as intervals of 100 milliseconds (100ms per observation), prices are determined as the mean of the interval and volume is aggregated. Resulting data are normalized with standard normal for better convergence and consistent results:

$$X_i = \frac{X - \mu_X}{\sigma_X}$$



Abnormality Detection with Hierarchical Clustering

We chose to use hierarchical clustering because it doesn't require specifying the number of clusters beforehand. Additionally, the results can be represented as a dendrogram (see Figure 2), which provides straightforward information to identify abnormal data points. Since stock data is more complicated, we decided to test different combinations of distance metrics and linkage criteria to identify the abnormal data and find the repeating abnormal data points.

Figure 2: Sample Dendrogram of Hierarchical Clustering. As a proof of concept, we created the first five tensors with values ranging from 0 to 11 and the sixth tensor with values ranging from 20 to 31. The dendrogram, generated using the Euclidean distance metric and complete linkage criteria, identifies and shows that tensor 6 is the abnormal data.

For distance metrics, we chose Euclidean (straight-line distance between points), Manhattan (sum of the absolute differences between coordinates), and cosine (measuring the cosine of the angle between vectors). For linkage criteria, we selected single (minimum distance between clusters), complete (maximum distance between clusters), and centroid (distance between the centroids of clusters)

Results & Discussion

Training TCN-AE

With around 120 epochs of mini-batch of size 32, the network was able to converge in around 120 epochs, more than the 10 epochs mentioned in the paper. We suspect it has to do with the significant increase in data length (T = 18000) compares to the 1024 long ECG data. Extended training to 10000 epoch did not seem to produce better distinct encoded data (Fig. 3, clustering result not shown).

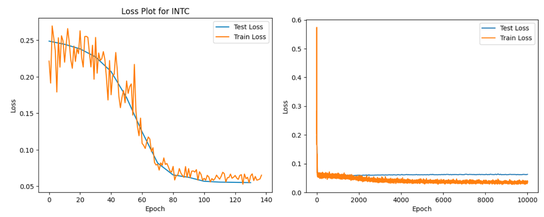


Figure 3. Sample training loss for TCN-AE training. Training loss reached a stable rate around 120 epoch. Test less were sampled every 10 epoch. Extended training with 10000 epochs does not seem to improve performance of the model (Test loss remain around constant).

Despite the low train and testing error, we believe there's room for improvement:

- **Change in loss function:** In the current implementation of TCN-AE and the one described in Thill et al, a logcosh loss function is used which is more tolerant to small changes. However, for our project a loss function which penalizes small errors more (comparatively) might perform better as we want our autoencoder to be sensitive to small changes.
- **More compact encoded layer:** Current implementation TCN-AE reduce data by factor of 50, which might still not be compact enough. A higher down sampling factor, or a fixed length encoded data length, despite possibly extending training time, might give better result in clustering performance.

Hierarchical Clustering

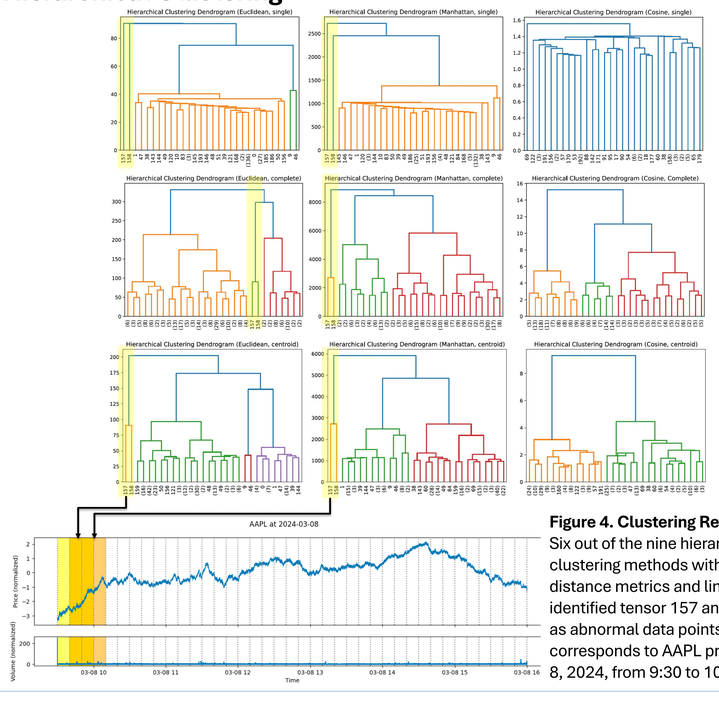


Figure 4. Clustering Results. Six out of the nine hierarchical clustering methods with different distance metrics and linkage criteria identified tensor 157 and tensor 158 as abnormal data points, which corresponds to AAPL price on March 8, 2024, from 9:30 to 10:10.

Acknowledgements

We would like to thank Professor David Byrd for his guidance and support throughout this project and the Financial Machine Learning course. We also want to thank Leopold Felix Spieler 24' for convincing Professor Byrd to allow us to work on the Double Deep-Q project before this research, which significantly contributed to the success of this research. Personally, Tom would also like to thank Mr. Tim Cook for announcing a stock buyback, which provide him a significant moral boost.

References

[1] S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling".
[2] M. Z. Rodriguez et al., "Clustering algorithms: A comparative approach," PLoS One, vol. 14, no. 1, p. e0210236, Jan. 2019, doi: 10.1371/journal.pone.0210236.
[3] F. J. Fabozzi, S. M. Focardi, and C. Jonas, "HIGH-FREQUENCY TRADING: METHODOLOGIES AND MARKET IMPACT".
[4] StatQuest: Hierarchical Clustering, (Jun. 20, 2017). Accessed: May 15, 2024. [Online Video]. Available: https://www.youtube.com/watch?v=7xHsRkOdVwo
[5] M. Thill, W. Konen, H. Wang, and T. Bäck, "Temporal convolutional autoencoder for unsupervised anomaly detection in time series," Applied Soft Computing, vol. 112, p. 107751, Nov. 2021, doi: 10.1016/j.asoc.2021.107751.
[6] L. Xie and S. Yu, "Unsupervised feature extraction with convolutional autoencoder with application to daily stock market prediction".
[7] W. Huang and P. Polak, "LOBSTER: Limit Order Book Reconstruction System," 2011. [Online]. Available: https://lobsterdata.com/. [Accessed: 15-May-2024].