Stock Price Prediction with Big Data Based on Machine Learning

Jingxuan Jiang

School of Sichuan Normal University, Chengdu 610101, China

862312940@qq.com

Keywords: Stock price prediction; big data; machine learning; Support Vector Machine (SVM); Long Short-term Memory (LSTM); Recurrent Neural Networks (RNN).

Abstract: Prediction of a stock price is an extremely challenging and complex task because of the dynamic economic environment. However, in this era of data, the availability of data and techniques makes it easier to do the prediction. As data mining techniques have been acquainted with and applied for financial prediction, in this paper, outlier detection and clustering which are two primary steps of data mining will be introduced to show how to process the data. Moreover, machine learning is ideal for stock prediction by analyzing historical data and giving the predicted result. This paper also introduces the detailed principles and technologies of the Support Vector Machine (SVM), Long Short-term Memory (LSTM), and Recurrent Neural Networks (RNN).

1. Introduction

In the era of big data, data mining techniques need to be applied to handle a large amount of data continuously acquired for a variety of purposes, besides, to improve their effectiveness, applications of machine learning to the prediction of the stock market has achieved good results in solving the problem of a large number of datasets.

1.1 Big data - Data mining

Data mining is a methodology for retrieving information from large-scale databases. It is a huge challenge to analyze this growing data in static or in dynamic form. Most traditional data analytic approaches and tools can't support the "big" scale. The stock price will be affected by various factors such as changes in economic conditions and political environment, which make it more complex to predict the stock price. To overcome the complicacy of the stock market with big data, outlier detection and clustering are introduced in this paper.

When dealing with very large data sets, automated tools are needed to find patterns and relationships. In large-scale data sets, one of the most vital things involved is to find outliers. This is defined as a sample or event that is very contradictory to the rest of the data set. (MichaelS., 2018) The change of the external environment will result in high-dimensional data sets. As a result, there are several anomalies in share prices which are usually considered to be an error or noise but are likely to carry valuable information. Where outliers exist, extra care should be taken to make sure that the stock price estimators used are robust. (Das, K., 2020)

Another technique of data mining that can be used in stock price prediction is clustering. One of the techniques of Clustering - K-means clustering is the process of grouping a collection of objects into classes of similar objects, a method for automatically classifying a series of patterns into clusters for the similarity-based introduction. Due to frequent changes in the stock market, the technique is required to be adaptable and can handle vast amounts of data. K-means clustering is scalable to a huge data set and also faster to large datasets and can adapt the new examples very quickly. It is also very simple to implement and can generalize clusters for different shapes and sizes.

1.2 Machine learning

In the 1950s and 1960s, after electronic computers came into use, algorithms for modeling and analyzing high volumes of data were soon developed. From the very start, 3 main areas of machine

learning have evolved. Hunt et al. (1966), Nilsson (1965), and Rosenblatt (1962) respectively explained the classic research on symbolic learning. Over the years, sophisticated approaches have been advanced in the 3 fields (Michie et al., 1994): k-nearest neighbors, Bayesian classifiers, and discriminant analysis and other statistical or pattern recognition methodologies; inductive learning of symbolic rules, such as top-down induction of decision trees, induction of logic programs and decision rules; artificial neural networks, such as the multilayer feedforward neural network with backpropagation learning, the Hopfield's associative memory, and the Kohonen's self-organizing network. (Kononenko, I., 2001)

As machine learning evolves, it has been applied in many research fields especially in finance and economics. For example, machine learning allows to transferring of algorithmic trading into smart dealing. machine learning is able to analyze the historical market patterns and come up with the best marketing strategy, making the transaction prediction more accurate.

2. Data mining and K-means clustering

Data mining is highly effective for stock prediction, which is used to analyze observation data sets (usually large ones) to explore unexpected connections and generalizing data in new modes. The connections and summaries obtained through a data mining implementation are commonly described as models. (Hand, D. J., 2007). Data mining contains the following methods: outlier detection, clustering, classification, tracking patterns, association, regression, and prediction.

Clustering is a widely accepted unsupervised learning technology for data mining, which is a method to locate similar data objects into clusters on the basis of some similarities.

K-means is amongst the most simplistic unsupervised learning algorithms that have a wide range of uses in the fields of image segmentation and information retrieval. (Wagstaff, K., Cardie, C., Rogers, S., & Schroedl, S., 2001).

The objective function is:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} W_{ik} \|x^i - \mu_k\|^2 \quad (1)$$

Where W_{ik}=1 for data point x^i if it belongs to cluster k; otherwise, W_{ik}=0. Also, μ_k is the centroid of x^i's cluster.

The fundamental algorithm can be described as below:

(1) Initialize the parameters to the existing model.

(2) Presuming that the present model is valid, cluster membership for data items is determined.

(3) Suppose that the data membership relationships acquired in step 2 are true, the present model's parameters should be re-estimated to generate a new version.

(4) This ends if the present model and the new one are close enough to one another, otherwise back to step 2. (Bradley, P. S., & Fayyad, U. M., 1998).

3. Machine learning for stock prediction

Machine learning is a significant technology in the stock prediction field. Many researchers have used Machine Learning technology to forecast stock prices or evaluate the predictive relationships between variables etc. David Enke (Missouri University of Science and Technology) and Suraphan Thawornwong use machine learning to assess the forecasting relationships between many economic and financial variables in data mining. (David Enke, Suraphan Thawornwong, 2005). Saahil Madge applied a Support Vector Machine to predict share prices. He computed the price variability and momentum of the individual shares and the industry as a whole with the daily closing prices of 34 technology shares. The great success of machine learning over the years has changed the way researchers use information, data, and optimal analysis and predictions.

3.1 Long Short-term Memory (LSTM)

Since the data used in stock price forecast is continuous data that need to be processed in a specific order, it is necessary to clearly understand them. Long Short-term Memory is a special

Recurrent Neural Network structure, which has the powerful function of preserving sequence information and has achieved good results in various sequence modeling tasks. In contrast to a standard feed-forward neural network, there are feedback linkages in Long short-term memory. It can handle not only individual data points (images etc.) but also the entire data sequences. (stock price, financial instruments, speech and, video, etc.). (Kai Sheng Tai, Richard Socher, Christopher D. Manning, 2015).

3.2 Support Vector Machine (SVM)

Support Vector Machine is another technology that can be applied to stock prediction. It is a supervised learning model with relevant learning algorithms that analyze data for the purpose of classification and regression analysis. Support Vector Machine is a computer algorithm that learns from examples and assigns labels to objects. Support Vector Machine works well with a clear margin of separation and with high dimensional space. Apart from executing linear classification, Support Vector Machine is also capable of carrying out non-linear classification efficiently with the so-called kernel trick that explicitly maps the input to high dimensional characteristic spaces. (Cortes, C., & Vapnik, V., 1995) Thus, the Support Vector Machine is highly effective in stock prediction.

3.3 Recurrent Neural Network (RNN)

Recurrent Neural Network is an artificial neural network in which the connections between the nodes form a dimensional graph along the time series. Each input is based on the previous input. To handle this kind of data, Recurrent Neural Network can be introduced for the reason that they are designed to learn sequential or time-varying patterns. Recurrent Neural Network is superior to other traditional modes of methods when it comes to learning the non-linear characteristics of a sequence for the reason that it is memorability, parameter sharing, and turning completeness.

The traditional neural network does not consider any sequence when processing the input and proceeding to the next step. Feed-forward networks cannot understand sequences because each input is considered independent. And the information in a feed-forward neural network can only move in one way from the input layer to the hidden layer and then to the output layer. The feeds information passes and only touches a given node once.

However, the working process of Recurrent Neural Network eclipses the traditional one in dealing with sequential data. Compared with the traditional neural network, Recurrent Neural Network comprises an input layer, a hidden layer, and an output layer working in a standard sequence, which can process sequential data effectively. The input layer fetches the data and sends the filtered data to the hidden layer. The hidden layer includes algorithms, neural networks, and activation functions that extract useful information from the data. Ultimately, the information is delivered to the output layer to give the intended result. (Das, K. 2020).

The process of how Recurrent Neural Network deal with input data can be explained by the following schema Figure 1:

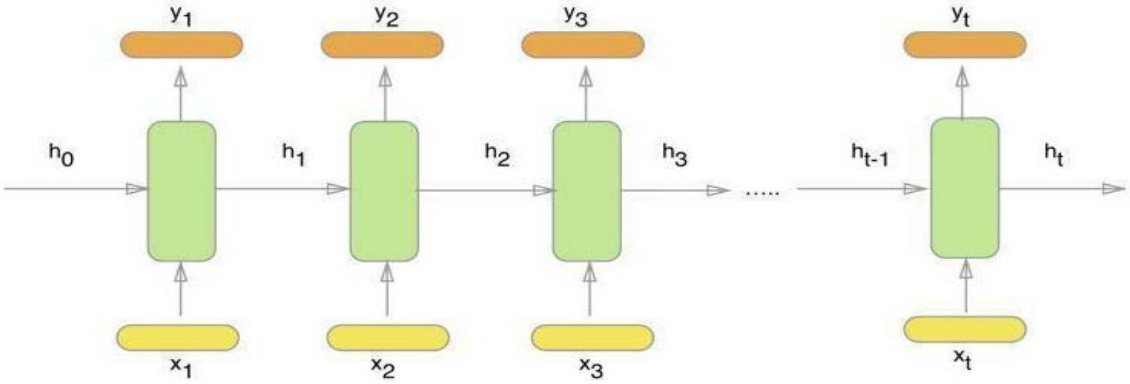

Figure 1. Working process of Recurrent Neural Networks (Kostadinov, S., 2017)

Here x_1, x_2, x_3, ..., x_t represent the input data, y_1, y_2, y_3, ..., y_t are the predicted data, and h_0, h_2, h_3, ..., h_t save the information about previously entered data.

The information passing through the architecture goes through a loop. To make decisions, Recurrent Neural Network will take into account both current input and previous input, and it allocates the same weights and biases to each layer of the network. As a result, all independent variables are transformed into dependent variables. And it is the loops in the Recurrent Neural Network that ensure the information is stored in the LSTM memory.

There are four categories of Recurrent Neural Network and the classification basis is the relationship between the number of inputs and the number of outputs. And the four type includes one to one, one to many, many to one, many to many. Each type is suitable in different situations, for example, a typical application for one to many is music generation and many-to-one can be used in sentiment analysis. (Das, K. 2020)

4. Stock price prediction with big data based on machine learning

To predict the future stock price by using historical data with big data technology and machine learning, the method focuses on the analysis of trends of stocks' price such as daily opening, high, low, and closing prices. In addition, other features that may be considered are used in prediction for increasing accuracy in the prediction, for example, volume and relative strength index, etc.

Some general steps are as follows:

(1) Using data mining of big data to extract unknown information from large datasets help researchers to focus on the most important information in data repositories.

(2) Using clusters to classify a large body of data into smaller groups that share the same characteristics.

(3) Choosing an appropriate network model where some of them are shown in section 3.

(4) Specify a network topology.

(5) A predicted outcome will be generated.

(6) There will be an error if the result is compared with the expected value.

(7) Variables will be adjusted if errors are propagated through the same path.

5. Conclusion

In this paper, several techniques have been illustrated to predict the stock price. To conclude, the use of clustering is touted as efficacious by many methods and K-means clustering is one of the most widely used ones, which minimizes within-cluster variances in the forecast. In addition, machine learning can be applied to the forecast by analyzing past data. This article also detailed the method of Support Vector Machine, Long Short-term Memory, and Recurrent Neural Networks. Since the data used in stock price prediction is time series data and Long Short-term Memory networks are able to classify, process, and make predictions based on such kind of data, as lags of uncertain duration between critical events exist in a time sequence. Support Vector Machine is also introduced as a technique to be applied to prediction because it has a good separation effect and is able to realize high-dimensional space. Recurrent Neural Network can also be applied to stock price prediction with its internal memory to process variable-length sequences of input data of stock price, making the output prediction come up to be a more accurate one.

References

[1] Michaels. (2018, January 1). Inside HPC. Retrieved from https://insidehpc.com/2018/02/ outliers-why-so-important/

[2] Das, K. (2020, November 5). Dataaspirant. Retrieved from https://dataaspirant.com/how-recurrent- neural-network-rnn-works/

[3] Michie D., Spiegelhalter D.J., Taylor C.C (eds.) Machine learning, neural and statistical classification, Ellis Horwood, 1994.

[4] Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in medicine, 23 (1), 89-109.

[5] Hand, D. J. (2007). Principles of data mining. Drug safety, 30 (7), 621-622).

[6] Wagstaff, K., Cardie, C., Rogers, S., & Schroedl, S. (2001, June). Constrained k-means clustering with background knowledge. In Icml (Vol. 1, pp. 577-584).

[7] Bradley, P. S., & Fayyad, U. M. (1998, July). Refining initial points for k-means clustering. In ICML (Vol. 98, pp. 91-99).

[8] Enke, D., & Thawornwong, S. (2005). The Use of Data Mining and Neural Networks for Forecasting Stock Market Returns. Expert Systems with Applications, 29 (4).

[9] Kai Sheng Tai, Richard Socher, Christopher D. Manning (2015). Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. ArXiv, 1503.00075.

[10] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20 (3), 273-297.

[11] Kostadinov, S. (2017, December 2). Towards Data Science. Retrieved from https://towardsdatascience.com/learn-how-recurrent-neural-networks-work-84e975feaaf7?gi=a8c0d229054c