A New Approach Using Weibo Data to Predict the China Shanghai Stock Market

T. Xu, H. Zhang
Information School
Capital University of Economics and Business
Beijing, China

***Abstract*—In recent years, researches using big data of SNS to predict trends are studied extensively these days. Weibo, the most famous Chinese SNS, is playing an important role for people getting and sharing information. In this paper, a large amount of text data were scraped from Weibo by web crawler, and then were used to build a model to forecast the trend of Shanghai securities composite index. The model based on Clustering and Neural Networks can help the investors to make better decision for their investment. The method proposed here also can provide import hint for related studies such as data mining..**

Keywords-clustering; data acquisition; Shanghai composite index forecasting

I. INTRODUCTION

Big data mining which is based on data generated by foreign social networks mainly used to conduct research on Twitter, including Twitter hot topic trend forecasting, and prediction of box office revenues by using of the contents of tweets and the results of the presidential elections, etc[1,2,3,4]. This article attempts to use social network data to make predictions about the domestic stock market[5]. Daily closing price of Shanghai Composite Index was used to forecast the trend of stock market and this is because Shanghai Composite Index to a large extent reflects the overall trend in the domestic stock market[6]. A web crawler was made to collect enough data from Sina Weibo. Use Python, BeautifulSoup, MySQL and other tools to clean and quantify crawled data, select and quantify the relevant dimensions of the stock movements, then, with the use of artificial neural network algorithm, a forecasting model was built on Shanghai Composite Index closing price. Extracted Sina Weibo Contents were used to forecast the prices of Shanghai Composite Index, at last verify the effect of prediction by using samples to test the model[7,8]. The results show that Sina Weibo articles associated with Shanghai Composite Index, whose bullish and bearish stock market conditions has a strong correlation with the trends of Stock market especially when the data set which is for prediction has distinguishing characteristics, prediction can be more accurate. This result may to some extent help the stock market investors to make better decisions.

II. SINA WEIBO DATA ACQUISITION AND CLEANING

Data acquisition strategy is to make use of a web crawler crawling and Sina Weibo API to gain access to data. The amount of microblogs where data were extracted from is about 30 billion, effectively crawled text records are approximately 60,000, and the amount of data files in total is 6993.

In this article, cleaning tools for raw data were chosen, including advanced programming language Python, web crawler curl, HTML parsing software BeautifulSoup and MySQL. By using a unique model of regular expressions and cleaning the crawled 6,993 HTML files, a total of 6993 microblogs were obtained. Data which has been cleaned is the data related to the Shanghai Composite Index data accumulated for 950 days from January 4, 2010 to December 27, 2012. Sina Weibo users' emotional data for Shanghai Composite Index has been stored into three dimensions, including positive emotions, cautious mood and negative emotions. To facilitate the establishment of forecasting models, data cleaning and word frequency calculation on mainly-analyzed dimensions should be processed at the same time[9].

III. BUILDING PREDICTION MODEL AND TESTING

This paper innovatively combines K-means clustering algorithm and neural network algorithm, extracting their advantages and personalizing the original basis, to further neural network algorithm with variable data sets and self-correction features. Depending on the input dimensions including the mood of Sina micro-blog users, amplitude and frequency of mood changes of the user, the paper intends to design a predictive model.

A. Algorithm Analysis

1) Clustering: In this study, the clustering algorithm is mainly used to classify data collection, including training data set and testing data set. According to the clustering results, that is to determine data across the range selected in the process of the model building and the model prediction. Depending on the difference of clustering results, the data set across the range for the model building and the model prediction will have a corresponding change. So in this paper, the data set used to predict is variable, and then increases the variable dataset characteristics for the prediction model. Compared with the range of strategies of taking fixed data, dynamic variable data set can further improve the prediction accuracy of the model.

2) Algorithm processing: This paper selects K-means clustering algorithm to handle the data. In the clustering process, there is a certain transition region between the various types. Therefore, this paper uses fuzzy clustering and further classifies the transitional data set to improve the model results. There is an interaction between the change of bearish factors and the change of Shanghai Composite Index closing price. So

the data in the transition region between different types cannot be absolutely divided in a certain concreting class. In view of this, fuzzy clustering mechanism, assuming the data items in the transition region belong to all classes near simultaneously, in the actual clustering process, according to attribute values, computes different classes of the right weight, and according to forecast demand, dynamically determines which class to belong. For further analysis, the cluster size parameter should be added in this article for a good class having been divided, and allowing nesting occurs between the classes according to the changes in the data set spans.

3) Variable data set: This article uses the Sina micro-blog information associated with the Shanghai Composite comment to predict the future trend of the Shanghai Composite Index. A key issue to be addressed is the prediction data set to determine the time span, that is to say for a point in time t, if we are to predict the ups and downs and the trend duration of the Shanghai Composite Index after a point in time t, there is a must to consider how much time span dating back before the time point t. Assuming the time span Δt, so the data set used to predict is located [t-Δt, t] . As shown in To the Shanghai Composite Index forecast for the market trend, if the time required for the use of data sets span is set to a fixed value, you may receive the following situation:.

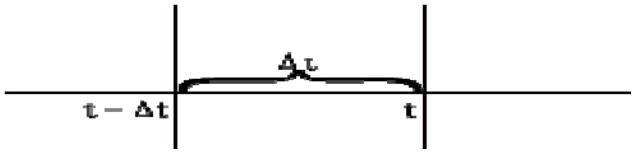

FIGURE I. TIME SPAN

To the Shanghai Composite Index forecast for the market trend, if the time required for the use of data sets span is set to a fixed value, you may receive the following situation:

(1). Data in a time span in the clustering process becomes into a single class, clearly different from other classes.

(2). Data in a time span in the clustering process, due to the less obvious features, you need to take fuzzy clustering strategy to determine the specific categories after further calculations.

(3). Data in a time span in the clustering process comprises a plurality of classes, which can be clearly separated.

So, when the time is set to a fixed value, only the case of the first is suit for the requirements of this article, and for the second and third cases, if directly used to predict, predict results of different types of data may be different or even opposite, which will affect the prediction accuracy. In view of the above circumstances, for setting the time span Δt, dynamic strategy is adopted herein, values are not fixed Δt from the time t back forward Δt time period, until all data objects within the time span are divided into two or more classes that have significant differences. The prediction data set can use data objects closest the time t, thereby it can ensure that the data set is used to predict the data objects, having as many common, which to some extent improve prediction effect model.

B. The Selection and Improvement Prediction Algorithm

Because of the nonlinear characteristics between the Shanghai Composite price and Sina micro-blog user sentiment, the paper select error back pass neural network (Back propagation), referred to as the BP neural network [10]. In this paper, the main use of the prediction model are the K-means clustering algorithm and back propagation artificial neural network classification algorithm and the appropriate improvements on the basis of the two are made. The study mainly includes the two following points:

(1). During the process of clustering and neural network combined with the clustering process, the class K on behalf of the number of arguments will serve as the input layer neurons neural networks, K value is finalized by the neural network.

(2). By neural network algorithm, we will get the prediction model eventually. In the statistical process model, the neural network will remember all of the test results, and predict the results of the updated history by the use of newly acquired knowledge before the predicted results take effect, improving overall accuracy of model predictions.

C. Modelling

This section will focus on the use of specific algorithms to select the previous sections and combine with both Sina micro-blog data and the Shanghai Composite closing price data to build predictive models.

1) Process analysis: The prediction model in this paper includes cluster-led training module mainly to the K-means clustering algorithm and classification module driven by the back-propagation neural network algorithm. Clustering and classification module are interrelated by input layer, since data object properties selected in clustering module is corresponding to the neurons node of the classification module input layer, and the modify of right weight of each node of the neural network module will also affect the calculation of the distance in the clustering module. The overall framework of the model is shown in Figure II..

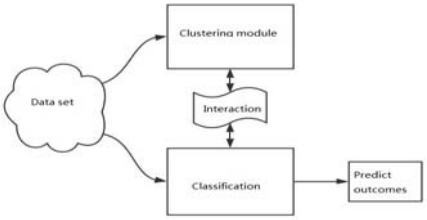

FIGURE II.THE OVERALL FRAMEWORK OF THE MODEL

Clustering module and classification module influence each other through interactive modules, clustering module mainly reads the data from the training data set. According to the data value of different heavy object properties, we use K-means clustering method to produce K classes. We use the interval to determine the training set according to the categories of current time, the result is passed to the interactive module. The process of clustering module is shown in Figure.

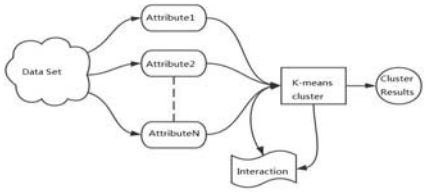

FIGURE III. THE PROCESS OF CLUSTERING MODULE

Classification module is shown in Figure IV..

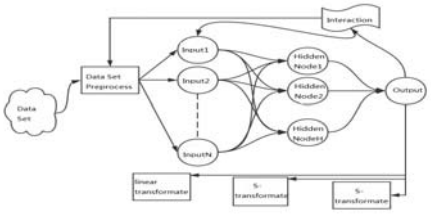

FIGURE IV.CLASSIFICATION MODULE

In classification process, BP neural network obtains classification results from the previous step in the interactive modules, and takes K value as input layer neurons nodes. Simultaneously, according to the data sets span parameters in the interaction module, we determine the useful range of the data set, and then after dataset pre-processing module, we take the interval of the data set as a whole object, extract the corresponding attributes and input to the neural network as the source of the data. The results of output node will function according to the error back propagation and modify the corresponding value of each node from the input layer to the hidden layer and from the hidden layer to the output layer. And adjust the value of K, the transfer it to the interaction module between classification and clustering module. The main role of interaction module is to provide transmission parameters between classification and clustering module, including the parameter K used in cluster and the parameter T used for classification (T represents the time span of the training set).

2) Model: This model is mainly complied by implement Octave, and the data dimensions used in this model are shown in Table I..

TABLE I. THE PROPERTIES OF TRAINING SET

Code	Variable name	Explanation
m1	mean_value_good	The average value corresponding to the bullish sentiment
v1	velocity_good	The speed of the bullish sentiment change
m2	mean_value_bad	The average value corresponding to the bearish sentiment
v2	velocity_bad	The speed of the bearish sentiment change
f	mood_frequency	The frequency of bearish sentiment change
t	time_cover	Time span, corresponding to the number of the training set used to predict
v6	stock_rate	The Shanghai Composite Index closing price rate of change
r	result	prediction reference value

By using two years of data as the training set, the model results for the above parameters of nonlinear polynomial function:

$$r = (\sqrt{m1^2 * 8032 * \pi} - m2 * 25) * 0.324 / f - (v1 * 0.42 + v2 * 0.66) * 0.25 - t * v6 * 0.52$$

D. *Model Testing*

The experimental data is collected within total 3 years from January 4, 2010 to December 26, 2012, we use the data of the first two years to train model and use the data of the third year to test the model.Something can be found from the figures. When the index parameter has obvious characteristics, the predicted effect is more obvious.Figure V. and Figure VI. have shown the entirety of 2010 and 2011 respectively.The model is used to examine the data set. The predicted results of the model are shown in Figure .

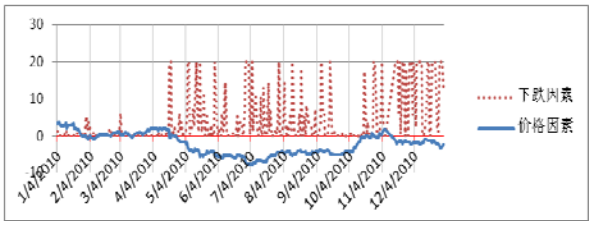

FIGURE V.THE PREDICT PROGRESS MODEL OF 2010

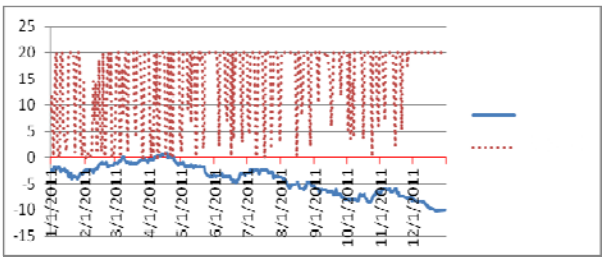

FIGURE VI.THE PREDICT PROGRESS MODEL OF 2011

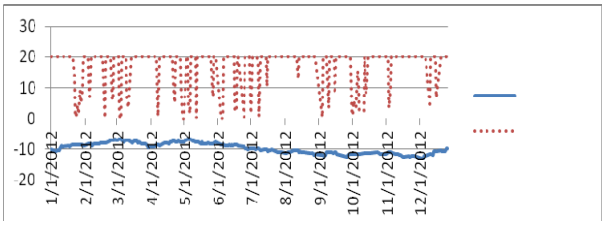

FIGURE VII. THE PREDICT PROGRESS MODEL OF 2012

The result of testing model is shown in Table II.

TABLE II. THE STATISTICS OF ACCURACY OF THE MODEL

Conditions of predicting data set	Precision of prediction	Accuracy
classification is obvious	high	87.2%
classification is obvious relatively	Common	62.4%
classification is not obvious	Low	25.2%

E. Summary

This model, based on the characteristics of the property of the research data, proposed variable data set intervals idea that dynamically determined the time span for training and prediction data sets. The model mainly uses the back propagation artificial neural network algorithm, set the weight of each attribute by back propagation, and in order to improve the prediction accuracy of the model, the study introduces automatic correction of thinking, that is to say, before the model predictions take effect, it modifies the pre-predicted values based on neural network new experiences and knowledge.

In this paper, the model consists of two parts, the clustering and classification section, the two parameters transfer data through data interaction module and modify for each other. During operation, clustering module extracts data from the source object of study, and re-clusters the data based on the initial weight property characteristics, determines the interval data set for prediction. Data collection interval parameter is passed to the classification module through interactive modules and classification module uses back-propagation neural network algorithm to determine the characteristics of the data attribute weights based on the training set. The weights and K values are fed back to the clustering module through interactive modules, and then optimize the clustering results. In the study, two-thirds of the data is used as the training set, and the remaining for testing. Test results show that, under the obvious characteristics of the training set, the prediction precision which can reach 87.2% is relatively high.

IV. CONCLUSION

Firstly, the study analyzes and compares the current research on social networks and then select the country's most influential current social media-Sina micro-blog-for the study. Finally, we analyze the trend of the ups and downs of the stock market Shanghai Composite index and establish the prediction model.

According to the prediction model, this study draws the following conclusions:

Based on the content of Sina micro-blog related to the Shanghai Composite, to a certain extent, you can predict the future trend of the Shanghai Composite Index.

In the content of Sina micro-blog, the greater the bearish sentiment is, the greater the likelihood of the Shanghai Composite Index fell is, while the larger the bullish sentiment is, the less likely the Shanghai Composite Index fell.

Validity of the model prediction, i.e. at some point in time, making a future prediction according to the data set will last how long time depends on the time span of the variable data sets, generally proportional to their relationship, and predictive validity will be dynamically adjusted because the model self-correcting feature.

Law previously mentioned is the basic law; the most important is the prediction accuracy can reach 87.2% when the data set has the obvious characteristics. Since the time this data selected is about three years, a relatively short time, it can be extended to a longer data set for testing in future studies. In addition, this study did not consider the transaction fee issue, the fee will be considered included in the prediction model next step.

ACKNOWLEDGMENT

This research was supported by Funding Project for Research Level's Improvement by CUEB and Research Project of CUEB(2014XJG022). It was also supported by Level of Scientific Research Improving Project of Beijing Municipal Education Commission 2014 and Beijing philosophy and social science project (13SHB015), and the Funding Project for Research Level's Improvement by BMEC.

REFERENCES

[1] Johan Bollen, Huina Mao, Xiao-Jun Zeng, Twitter mood predicts the stock market, NSF Grant BCS, 2010, 10: 1~8

[2] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe, Predicting elections with Twitter: What 140 characters reveal about political sentiment, International AAAI Conference on Weblogs and social Media Washington DC, 2010

[3] Sitaram Asur, Bernardo A. Huberman, Predicting the future with social media, Computing Research Association for the CIFellows Project, 2010, 3

[4] Daniel Gayo-Avello, A balanced survey on election prediction using Twitter data, arXiv, 2012, 4: 1~13

[5] Xiaobing Xiong, Gang Zhou, Yongzhong Huang, Jun Ma, Study on the technology of Sina Weibo topic popularity prediction, Journal of Information Engineering University, 2012, 13(4), 496~502

[6] Zhiyuan Liu, Xinxiong Chen, Maosong Sun, Mining the interests of Chinese microbloggers via keyword extraction, Front. Comput. Sci, 2012, 6(1), 76~87

[7] Li He, Yue He, Yeqing Huo, Weibo User Characteristics Analysis and Core Users Mining, Information system, 2010, 11: 121~125

[8] Ye Tian, Sina Weibo-Based Event Trend Analysis and Prediction Analysis, Wuhan University, 2012, 5

[9] Zhang Hongcheng, An Advanced Data Capture Method Based on Sina Weibo, Information and Computer, 2013, 11,54~55

[10] Junfei Qiao, Honggui Han, Analysis and Design of Feedforward Neural Network, Science Press, 2013, 1: 56~215