International Journal of u- and e- Service, Science and Technology
Vol.8, No.2 (2015), pp.35-44
http://dx.doi.org/10.14257/ijunesst.2015.8.2.04

Text Categorization System for Stock Prediction

Bozhao Li, Na Chen, Jing Wen, Xuebo Jin and Yan Shi

School of Computer and Information Engineering, Beijing Technology and Business University, Beijing, 100048, China
youhusky@gmail.com

Abstract

Due to the complex market environment, it is very difficult to get the accurate predict of the stake only by the data analysis method. This paper uses the text categorization method to predict the trend of the stock. We divide the text categorization method into the following three steps: Text representation, Feature selection and Text Categorization. By comparing several categorization methods including feature selections and feature spaces, etc., the results show that the SVM method with Information Gain and 1000 feature spaces can get the better performance for the predict of the stock with the news.

Keywords: *Text categorization, SVM, Information Gain, Stock prediction*

1. Introduction

Nowadays during increasingly developed technology of the World Wide Web and Internet, the data is becoming extremely rich. With the application of data recognition process, the information extracted from data has become the most important part in some areas of society, management field, finance and markets, etc. It is necessary to develop the valid method to understand the knowledge of the data.

In order to use these data and abstract the useful information, there have been a number of methods to mine data and study the various categories of information. However, most the data mining method are only for structured data. Because of the texts on the website from different fields like the expression, written, and output, the classical data mining technology cannot fulfill effectively for such data. Some text categorization methods can deal with the unstructured data, therefore it has been a useful method for the extracting information from the unstructured data.

For example in the prediction of the stock, the news text information is a kind of unstructured data, which sometimes there has been a strong correlation with the trend. Researchers used the Mood Tracking Tool to analyze twitter over several million information and found that the twitter "calm" level can predict the Dow Jones industrial average trend index [1]. They assumed that public and financial experts present the same attention to the Dow Jones industrial average index, so their emotional expression will directly affect their investment decisions, thus these decisions will have an impact on the stock market, but this research needs further study [2].

Therefore, it is possible to develop network text processing technology to improve the prediction precision and get more accurate decision. We assure that the text mining technology will have more and more prominent role in today's monetary field like stock markets.

2. Text Representation

2.1. Related Works

Text categorization is defined that the process of text contents is assigned to the given certain or several classes. Before the 90's of the twentieth century, the dominant text categorization method is built on knowledge engineering which needs skilled personnel

ISSN: 2005-4246 IJUNESST

manual classification. As we know that artificial classification is a very time-consuming working and inefficiency. After 90 of the twentieth century, many statistics and machine learning methods are used in text categorization which has aroused great interest of researchers. At present, researchers have also started the research and get preliminary application of Chinese text categorization like information retrieval, digital library, automatic summarization and categorization of news group, text filtering, semantic analysis of words and document organization and management and other fields.

This paper develop a Text Categorization System with the following four stages as the text preprocessing, feature representation, feature selection methods and text categorization methods[4]. The text categorization system block diagram and programmable flowchart are shown in Figure 1.

Figure 1. System Block Diagram and Programmable Flowchart

2.2. Text Preprocessing

Before representing the text, the news text should be preprocessed because we need to unify the format on original data, which it is convenient for subsequent processing. Be compared with the English text classification, an important difference in Chinese text classifications is preprocessing stage: Chinese text needs to word segmentation, which is unlike English text words with spaces to distinguish. Because in Chinese, the useful term is composed by the single word and we have to extract the term from the whole article with so much words and without the distinct spaces between words, just like the English.

Therefore the Chinese text preprocessing includes segmentation and the removal of stop word lists. In information retrieval, in order to save storage space and improve searching efficiency, we need to filter out certain word that are called stop words automatically before processing of natural language data. These stop words are generated manually without any automation and will form a stop words list. In this research, we use ICTCLAS that includes 1208 words according to the Chinese stop word lists and authoritative English stop word lists [4].

2.3. Feature Representation

The text feature representation uses the feature configure on representation. In the process of representation, there are two key issues need to be considered: One is the choice of which features to characterize the semantic of the text, which means the text feature selection; The other is to choose which model can organize these features, which means the text representation model [5].

Different words for the document have different weights which can calculate the contribution of each word in the document. The weight calculation methods include Boolean Weighting, Term Frequency Weighting and TF-IDF (Term Frequency-Inverse Document Frequency weighting). Among these methods, Boolean Weighting has not taken into account the frequency of the word and Term Frequency Weighting has ignored the word is common or rare across all documents.

So here we select TF-IDF which considers all aspects to calculate the weight of the words, which is defined as $TF - IDF = TF * IDF$, where TF is Term Frequency and IDF is Inverse Document Frequency

A high weight of TF-IDF is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms.

3. Feature Selection Methods

Feature selection is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context.

Based on the feature we then compare the vector of unclassified document and vector of training set document. The most similar class of the document is selected and defined as the correct class in order to realize the classification process. The general feature selection methods are the following:

3.1. IG (Information Gain)

Information gain is the feature of the difference of information entropy before and after the text. Information entropy is a value to judge words whether or not belongs to the information provided in a class.

where $P(C_i)$ is the appearance probability of 'C_i' category, which means the number of category's documents divided by the number of total documents. $P(t)$ Is the appearance probability of 't' feature, which means the number of 't' feature

$$IG(T) = H(C) - H(C \mid T) \tag{1}$$

$$= -\sum_{i=1}^{n} P(C_i)\log_2 P(C_i \mid t) + P(t)\sum_{i=1}^{n} P(C_i \mid t)\log_2 P(C_i \mid t) + P(\bar{t})\sum_{i=1}^{n} P(C_i \mid \bar{t})\log_2 P(C_i \mid \bar{t})$$

divided by the number of total documents, and $P(C_i \mid t)$ is the appearance probability of 'C_i' category when 't' feature appears.

3.2. CHI Square Statistics

CHI statistics is a value to examine the degree of relation between a feature and a category. If 't' feature and 'C' category are independent, the statistics degree of 't' feature is zero.

The definition of CHI statistics about the 'C' category and 't' feature is the following: The simplify is

$$\chi^2(t,c) = \frac{[P(t,c)P(\bar{t},\bar{c}) - P(t,\bar{c})P(\bar{t},c)]^2}{P(t)P(\bar{t})P(c)P(\bar{c})} \tag{2}$$

If the $\chi^2(t,c)$ degree is high, the 'C' category and 't' feature are more

$$\chi^2(t,c) = P(t,c)P(\bar{t},\bar{c}) - P(t,\bar{c})P(\bar{t},c) \tag{3}$$

dependent.

3.3. Mutual Information (MI)

Mutual information indicates a correlation between two vectors in order to measure interdependence between two signals. In the text feature selection, if 't' feature is high

$$MI(t) = \sum_{t} p(C_i) \log \frac{p(t \mid C_i)}{p(t)} \quad (4)$$

appearance probability in 'C' class while is low in other classes, its mutual information will get higher. The calculation formula is as follows:

3.4. Expected Cross Entropy (ECE)

Expected Cross Entropy is a theory which only considers the feature appearance and not included in all situation. Although it will save time when we need to analyze large data, it reduces the preciseness.

$$CE(t) = p(t) \sum_{i} p(C_i \mid t) \log \frac{p(C_i \mid t)}{p(C_i)} \quad (5)$$

4. Text Categorization Methods

This paper will use two methods to select features and give the results that which is more suitable for the prediction by the text of the news.

4.1. K Nearest Neighbors (KNN)

Among several categorizations, KNN is one of the most classical method which is a non-parametric method used for classification. In the k-NN classification, the output is a class membership. An object is now classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If, $k = 1$ then the object is simply assigned to the class of that single nearest neighbor [6].

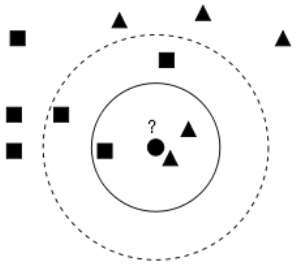

Figure 1. Example of *k*-NN classification

The test sample (circle) should be classified either to the first class of squares or to the second class of triangles. If $k = 3$ (solid line circle) it is assigned to the second class because there are 2 triangles and only 1 square inside the inner circle. If $k = 5$ (dashed line circle) it is assigned to the first class (3 squares vs. 2 triangles inside the outer circle).

The shortcoming of the k-NN algorithm not only is just sensitive to the local structure of the data but computationally intensive for large training sets. The curse of dimensionality in the k-NN context basically means that Euclidean distance is unhelpful in high dimensions because all vectors are almost equidistant to the search query vector (imagine multiple points lying more or less on a circle of with the query point at the center; the distance from the query to all data points in the search space is almost the same) [7-8].

4.2. Support Vector Machine (SVM)

Compare to the traditional classification method, support vector machine shows better performance in solving the small sample, nonlinear, high dimension space problem and especially in the text classification problem. Based on the outstanding performance, we select SVM method to solve the text classification issue of stock news.

SVM are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, a SVM training algorithm builds a model that assigns new instances into one category or the other, making it a non-probabilistic binary linear classifier. A SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. Further examples are then mapped into that same space and expected to belong to a category based on which side of the gap they fall on.

More formally, a support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be utilized to classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

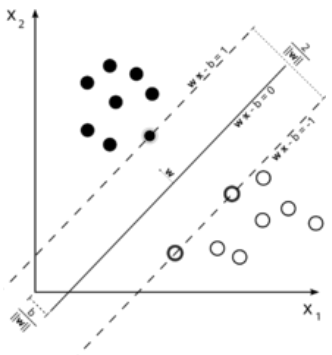

Figure 2. A Visual Principle about SVM

Classifying data is the common task in machine learning. Suppose some given data points each belongs to one of two classes, and the goal is to decide which class a new data point will be in. The data point is viewed as a p-dimensional vector (a list of p numbers), and we want to know whether we can separate such points with a $(p-1)$ dimensional hyperplane which is referred to as a linear classifier. There are many hyperplanes that might classify it. One reasonable choice as the best hyper plane is the one that represents the largest separation, or margin, between the two classes. So we choose the hyperplane so that the distance from it to the nearest data point on each side is maximized. If the hyperplane exists, it is recognized as the maximum-margin hyperplane and the linear classifier it defines is known as a maximum margin classifier; or equivalently, the perceptron of optimal stability [9].

5. Analysis and Result

5.1. Text Collections

In order to analyze how the alteration of stock influenced by news, we choose Poly Real estate Stock financial news for text analysis. Poly Real Estate Group Co, Ltd (SSE: 600048), is China Poly Group's large-scale real estate owned enterprises. The company is

located in Guangzhou and operate in Guangdong, Guangzhou, Beijing, Shanghai, Wuhan, Chongqing, Shenyang ten city real estate business.

We pick a 1000 Poly Real Estate (600048) news from the website, which includes prediction about next day's price is up (POS) or down (NEG). Each news is saved as TXT format and each of them are marked as positive or negative label of true reality. We take them as training set to train the model.

5.2. Evaluation Measures

In information retrieval contexts, precision and recall are defined in terms of a set of retrieved and a set of relevant documents [10].

In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the find:

$$Precision = \frac{\{relevant\} \cap \{retrieve\}}{\{retrieve\}}$$

For a text search on a set of documents precision is the number of correct results divided by the number of all returned results.

Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$Recall = \frac{\{relevant\} \cap \{retrieve\}}{\{relevant\}}$$

For text search on a set of documents recall is the number of correct results divided by the number of results that should have been returned.

In simple terms, high precision means that an algorithm returned substantially more relevant results than irrelevant, while high recall means that an algorithm returned most of the relevant results.

A measure that combines precision (P) and recall (R) is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score (F_1)

$$F_1 = \frac{2 * P * R}{P + R} \tag{6}$$

The F_1 can be interpreted as a weighted average of the precision and recall, here a F_1 reaches its best value at 1 and worst score at 0.

5.3. Analysis

The aim of this research is to select the most suitable method to achieve the stock predict [11]. We compare four parts of the project following different models, different selection features, different feature space and different parameter in SVM [12].

Firstly, we compare two methods about SVM and KNN by using Rapidminer [13]. RapidMiner is a software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics. The parameter of SVM is default that $C = 1, \gamma = 1$ .According to our text collections, the result is showed in Table 1.

Table 1. Comparison Result with Different Methods

Method	Recall	Precision	F1
SVM	0.83056	0.83403	0.832291
KNN	0.61099	0.67119	0.639677

Because of high F-score in SVM, we conclude from Table 1 that the SVM can obtain the better performance.

Secondly, we want to find the best parameter by using LIBSVM to analyze our model [14]. LIBSVM is a popular open source machine learning library developed at the National Taiwan University. The effectiveness of SVM depends on the selection of kernel, the kernel's parameters, and soft margin parameter C. A common choice is a Gaussian kernel, which has a single parameter γ. The best combination of C and γ is often selected by a grid search with exponentially growing sequences of C and γ, for example, $C \in \{2^{-5}, 2^{-3}, ..., 2^{13}, 2^{15}\}$ and $\gamma \in \{2^{-15}, 2^{-13}, ..., 2^{1}, 2^{3}\}$. The default parameter is $C = 1, \gamma = 1$.

Typically, each combination of parameter choices is checked using grid search by cross validation, and the parameters with best cross-validation accuracy are picked. The final model, which is used for testing and for classifying new data, is then trained on the whole training set using the selected parameters.

So in our model, according to grid search by LibSVM, we get the best parameter $C = 2^{11}, \gamma = 2^{-11}$ and, the result is showed in Table 2 and the best parameter is obviously competitive.

Table 2. Comparison with Different Parameters

Parameter	Recall	Precision	F1
$C = 2^{0}, \gamma = 2^{0}$	**0.71171**	**0.71289**	**0.712300**
$C = 2^{3}, \gamma = 2^{-3}$	**0.74267**	**0.74459**	**0.743629**
$C = 2^{5}, \gamma = 2^{-5}$	**0.77012**	**0.76967**	**0.769895**
$C = 2^{7}, \gamma = 2^{-7}$	**0.80982**	**0.80013**	**0.804946**
$C = 2^{9}, \gamma = 2^{-9}$	**0.82071**	**0.81974**	**0.820225**
$C = 2^{11}, \gamma = 2^{-11}$	**0.83056**	**0.83403**	**0.832291**

Thirdly, we use different feature selection methods to discuss: Information Gain (IG), Mutual Information (MI) Expected Cross Entropy (CE) and CHI square statistics (χ^2) [15] [16]. Then we choose the best selection which can improve the accuracy and speed into further analysis. The result is shown in Table 3 and we can see the information Gain shows the best result.

Table 3. Comparison with Different Features

Features Selection	Recall	Precision	F1
IG	0.83056	0.83403	0.832291
MI	0.73943	0.74767	0.743527
CE	0.77789	0.78081	0.779347
χ^2	0.80353	0.80417	0.80385
WE	0.76715	0.79138	0.779077

At last we discuss the relation between the feature spaces and the signal performance [17]. More space will analyze more features. But when we use our methods into Big Data, tremendous spaces is unavailable which we need to figure out the best score of space. The result is shown in Table 4.

Table 4. Comparison with Different Feature Space

Feature Space	Recall	Precision	F1
50	**0.60776**	**0.61147**	**0.609609**

100	**0.71448**	**0.73394**	**0.724079**
300	**0.74889**	**0.74737**	**0.748129**
500	**0.80353**	**0.80417**	**0.803850**
800	**0.81507**	**0.81316**	**0.814114**
1000	**0.83056**	**0.83403**	**0.832291**

By applying text classification system of the Poly Real Estate news in following these comparisons, we found that the SVM with Information Gain, 1000 feature spaces and the best parameter $C = 2^{11}, \gamma = 2^{-11}$ can get the best result and the accuracy of text has reached 83%, which mean that the stock is largely affected by financial news.

5. Conclusion

At present, only using data to predict the stock is limited and the influence of other impacts like news which can show the current situation is ignored. In our method, we use text categorization to predict the price of stocks. Based on the news of text categorization is proposed, which builds a structure of system to predict the stock. We also design and conclude some useful and valuable parameter to optimize our system, which obtains very good predictive performance.

Using text classification from the subjective and objective information will be a more sophisticated and mature method to extraction and classification. In the future, we will have a broader development prospects.

Acknowledgements

This work is partially supported by NSFC under Grant No. 61273002, 60971119 and the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions No. CIT&TCD201304025.

References

[1]. B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu and M. Demirbas, "Short text classification in twitter to improve information filtering", In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, ACM, **(2010)** July, pp. 841-842.
[2]. J. Bollen, H. N. Mao and X.J. Zeng, "Twitter mood predicts the stock market", Arxiv working paper, **(2010)**, Indiana University.
[3]. F. Sebastiani, "Machine learning in automated text categorization", ACM computing surveys (CSUR), vol. 34, no. 1, **(2002)**, pp. 1-47.
[4]. X. Li, and C. Zhang, "Research on enhancing the effectiveness of the Chinese text automatic categorization based on ICTCLAS segmentation method", In Software Engineering and Service Science (ICSESS), 2013 4th IEEE International Conference on, IEEE, **(2013)** May, pp. 267-270.
[5]. Uğuz and Harun, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm", Knowledge-Based Systems, vol. 24, no. 7, **(2011)**, pp. 1024-1032.
[6]. S. Jiang, G. Pang, M. Wu, and L. Kuang, "An improved nearest-neighbor algorithm for text categorization", Expert Systems with Applications, vol. 39, no. 1, **(2012)**, pp. 1503-1509.
[7]. R. Z. Xia, Y. Jia and H. Li. "A Text Categorization Method Based on SVM and Improved K-Means", Applied Mechanics and Materials, vol. 427, **(2013),** pp. 2449-2453.
[8]. N. S. Altman. "An introduction to kernel and nearest-neighbor nonparametric regression", The American Statistician, vol. 46, no. 3, **(1992)**, pp. 175–185.
[9]. L. H. Lee, C. H. Wan, R. Rajkumar and D. Isa, "An enhanced support vector machine classification framework by using Euclidean distance function for text document categorization", Applied Intelligence, vol. 37 no. 1, **(2012)**, pp. 80-99.
[10]. J. Davis and M. Goadrich. "The relationship between Precision-Recall and ROC curves", In Proceedings of the 23rd international conference on Machine learning, ACM, **(2006)** June, pp. 233-240.
[11]. S. W. Chan, and J. Franklin, "A text-based decision support system for financial sequence prediction", Decision Support Systems, vol. 52, no. 1, **(2011)**, pp. 189-198.

[12]. R. Luss and A. d'Aspremont, "Predicting abnormal returns from news using text classification", Quantitative Finance, (ahead-of-print), **(2012)**, pp. 1-14.
[13]. Manne, Suneetha, *et al*., "Features Selection Method for Automatic Text Categorization: A Comparative Study with WEKA and RapidMiner Tools", ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India-vol II, Springer International Publishing, **(2014)**.
[14]. C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines", ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, no. 3, **(2011),** pp. 27.
[15]. R. M. Balabin and S. V. Smirnov, "Variable selection in near-infrared spectroscopy: benchmarking of feature selection methods on biodiesel data", Analytica chimica acta, vol. 692, no. 1, **(2011)**, pp. 63.
[16]. T. Joachims, "Text categorization with support vector machines: Learning with many relevant features", Springer Berlin Heidelberg, **(1998)**.
[17]. J. Fan and J. Lv, "A selective overview of variable selection in high dimensional feature space", Statistica Sinica, vol. 20, no. 1, **(2010)**, pp. 101.

