Journal of Computational Information Systems5:6(2009) 1625-1631
Available at http://www.JofCI.org

One-class Classification based Finance News Story Recommendation

Zhongyu WEI†, Jun XUN, Xiaolong WANG

Intelligence Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China

Abstract

In this paper, we proposed an importance evaluation method for finance news story recommendation based on one-class classification. Based on the "important news stories" which are generated automatically as our training corpus, we used the one-class classification approach to evaluate the importance of each finance news story. This research not only quantifies the importance of each finance news story successfully, also makes it possible for ranking the results in finance specialty search engine in an innovative way. We investigated on the influence of features number and threshold to the performance of three one-class classification method Roccihio, k-means and one-class SVM. As experimental results shows, the method k-mean algorithm which had the best performance produced the precision up to 80% while maintaining recall at 95%.

Keywords: Information Recommendation; Financial Text Analysis; One-class Classification; One-class SVM; K-means

1. Introduction

People read finance news story to understand what is happening significantly to a company or stock. Generally, finance news stories are presented on online finance communities in portal websites (Hexun[10], Yahoo Finance[11] .etc) and searching websites (Google Finance[12]) based on manual editing and search engine respectively. On one hand, the advantage of editing the finance news story manually is of high accuracy and more target-oriented, however, it would results in narrowing the coverage areas. On the other hand, selecting the finance news story automatically using search engine based on key words matching technique can bring in the problem of low accuracy of the finance news story. Besides, it could hardly meet the customers' expectation without fully considering the relative importance of the news story stories when sorting them. For instance, news story "** celebrity appears at airport"(** 现身机场: http://ent.xinmin.cn/2009/10/14/2724506.html) was labeled as important news story of Shanghai airport in Google Finance only because it contains the key word "上海机场" (Shanghai airport). In this paper, we tried to solve the news recommendation problem in a novel way based on search engine, which can help users to filter important finance news stories.

If the news stories about one company or stock returned by search engine are treated as a single set, those of highly relative importance within the set can be regarded as normal points, otherwise, they are known to be outliers. Therefore, the determination on whether a news story is significant or not can be treated as a classification problem between normal points and outliers. It's obvious that the distribution for the normal points and outliers in the corresponding results about company and stock, returned by search engine, is not uniform, in which the news stories with high relative importance are dominated. Meanwhile, if we train the classification model based on manually labeling method, the model obtained may not be well applied in the general way, due to non-uniform distribution property of corpus labeled manually. "One-class classification can be used for classification problem where one of the classes is sampled very

† Corresponding author.
Email addresses: yutouwei@gmail.com (Zhongyu WEI)

December, 2009

well, while the other class is severely under-sampled"[1]. In our research, the one-class classification approach is applied to solve this critical problem on how to determine the significance of finance news stories when only positive data (important finance news stories) is available.

Related work. The most similar research to ours comes from V. Lavrenko etc. which implemented system Enalyst based on Bayes language model to accomplish the recommendation process [9]. Different from evaluating the importance degree of news story in our research, Enalyst labeled news story as different tendency such as surge, slight fall and so on by correlating the content of news stories with trends in financial time series, in order to help users to predict the forthcoming trends in stock prices. In this paper, finance news story recommendation is treated as a one-class classification problem.

One-class classification, also called outlier detection, is widely used in Information Security, data mining and so on. For instances, the outlier detection technologies were employed to detect network intrusion attempts [5], detect the outliers of price in stock market[6] and do document classification[4]. However, until now, this approach has not been introduced for importance evaluation in finance area.

According to the standard by TAX [1], there are three kinds of one-class classification methods based on density methods, boundary methods and reconstruction. With some adjustments to the above fundamental methods, [3] proposed a method based on distance; [1, 2] used SVM to solve one-class classification problem. In the next section, three one-class classification methods Roccihio, k-means and one-class SVM employed in our research are introduced.

2. One-class Classification Method

The major task of one-class classification is to make a description of the target class and detecting which (new) objects resemble this training set. In the training process, a threshold was computed based on a predetermined recall, while in the prediction process, each news story was evaluated by calculating the similarity between its vector and the model established in the training process, then it would be labeled as an important news story if the similarity was greater than the threshold. Here are some symbols or formulas used:

Cosine formula is used to compute the similarity between two vectors:

$$sim(d_1, d_2) = \frac{d_1 * d_2}{|d_1||d_2|} \tag{1}$$

Where d represents the a news story vector, each dimension corresponds to a feature pair consisting of one word and relative weight, where weight is valued by TF and Rel(d) represents the importance degree of d.

2.1. Roccihio

Roccihio is the classic method for document routing or filtering in information retrieval. Since it is easy and convenient to be understood and implemented, it is also used as a benchmark algorithm for classification. In this method, a prototype vector is built for target class C, and a news story vector d is evaluated by calculating the similarity between d and the prototype vector. The news story with a higher importance degree will be labeled as "important news story". The importance degree is calculated by the following formula:

$$\mathrm{Re}l(d) = Sim(d, M_c) \tag{2}$$

M_c represents the prototype vector.

The prototype vector M_c is computed by treating all the news stories in training corpus as one news story and transformed it to a vector.

2.2. K-means

As Rocchio do, k-means also try to model the target set by training corpus. However, instead of

establishing a prototype vector, in k-means method, it assumes that data is clustered and can be characterized by a few prototype vectors μ_k. In our research, we aimed to find out the model maximizing the total importance degree in training data. The algorithm started with k random center vectors. All the news stories in training data were then assigned to the most relevant category (represented by center vector). After that, the center vector was updated to the mean of this set of news. The estimation process repeats until the center vectors were converging (or set a predetermined number of iteration). In the k-means algorithm, the criterion for convergence of the center vectors was to maximize the following value:

$$\mathrm{Re}l_{total} = \sum_i \mathrm{Re}l(x_i) \tag{3}$$

Formula 4 is used to calculate the importance degree of d is bellow:

$$\mathrm{Re}l(d) = \max_k[Sim(d, \mu_k)] \tag{4}$$

μ_k represents center vector.

In the prediction process, we computed the importance degree for test news using formula 3 and labeled the news story whose score is above the threshold as "important news story".

2.3. One-class SVM

The more recent SVM algorithm has been more applied in two-class classification problem recently with good performance. Scholkopf and Smola proposed a modified SVM methodology for solving the one-class classification problem. Essentially, after transforming the feature via a kernel, they treated the origin point as the only member of the second class, and tried to find out the hyper plane separating the target set from origin. By using "relaxation parameters", they could control the portion of positive samples in the training set. In the prediction process, the standard two-class SVM techniques were employed, which is shown in Figure 1:

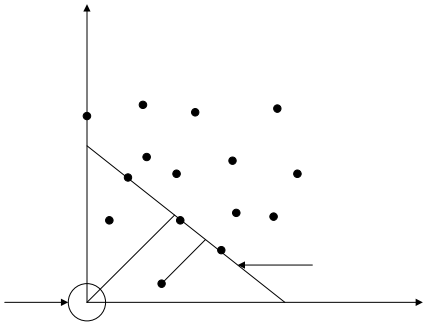

Fig. 1 One-class SVM

In order to judge whether an object belongs to target set or not, we should find out the function:

$$f_w(x) = (w * \varphi(x)) \tag{5}$$

Where $\varphi(x)$ represents kernel，w represents hyperplane. If the $f_w(x)$ is greater than threshold ρ, x is labeled as the target set. The following formula is used to find w and ρ：

$$\min_{w\in F, v\in R_+^N, \rho\in R}[\frac{1}{2}\| w \|^2 - v\rho + \frac{1}{N}\sum_{n=1}^{N}\varepsilon_n](w * \varphi(x)) \geq \rho - \varepsilon_n, n = 1, 2, ..., N) \tag{6}$$

Where ε_n represents the dot in space after transforming, v represents the "relaxation parameters".

In our research we used the LIBSVM(version2.89, 2009). This is an integrated tool for support vector classification and regression which can handle one-class SVM using the algorithm proposed by Sholkopf and Smola. We use POLY kernel and other parameters by default.

3. Experiments and Analysis

In this section, we employed these three one-class classification methods for obtaining the relative importance of the finance news stories, based on the pre-specific corpus. Apart from comparing the performance of the three approaches, we also investigate on the influence of the different parameters on the performance of each approach including the threshold and the number of features.

3.1. Corpus

The corpus for the experiment is mainly divided into two parts: the training dataset[13], and testing dataset[14]. The training dataset is created automatically, while the testing one is labeled manually.

For the training corpus, those news stories which are of potential for influencing the stock price are regarded as important ones. According to this, we generate our training corpus. First of all we define two terms here: critical time interval and significant day. Since the news available during the period between previous day's stock market closing and that day's market closing have the most significant and direct influence on the stock price of one company for that specific day, we define this period as the critical time interval for that day's stock price. Besides, if a stock is trading while the stock index falls, or the stock is limit while the stock index rises, we define that specific day as a significant day to this stock.

According to the stock exchange records provided by China Stock Exchange from March, 2008 to March 2009, we find out all the significant days for each A share stock in Shanghai and Shenzhen stock market, then collect all the relevant news published during the critical time interval corresponding to each significant day. All the news stories collected are used for training corpus. Similar approach for collecting corpus has been proposed in [8][9], however, the corpus obtained was applied to predict the price of the stock. Experiment result shows that, our method of generating corpus is able to stimulate the true distribution of "important news stories". At the meantime, we can adapt the change of focus in stock market by changing the date interval for collecting news stories.

For the testing corpus part, we labeled a collection of important news stories and irrelevant ones as well manually. The corpuses are labeled manually, and then only the intersection contents are kept for purpose of training. The distribution of the training corpus and testing corpus are shown in the Table 1.

Table 1 Corpus for Experiment

Corpus	Important	Outliers	Generate method
Training	4396	0	Automatic
Testing	1967	1984	Manual

3.2. Pre-processing s

We performed the extraction of features for training corpus under our supervision before carrying out experiments. The extraction process is described as follows: First, we counted the occurring frequency of each word in the training corpus. After that, we filtered the words to obtain the features wordlist in accord with the frequency of each word, which was based on a finance dictionary unpublished. The advantage of our approach is that it is easy to implement and also performs well when processing the corpuses from the area of finance. We have evaluated its performance with respect to the traditional TF-IDF approach, and it has shown that our approach has obvious advantages.

3.3. Evaluation Method

In our research, we used recall, precision, BEP and F1 to evaluate the performance of each methods.

Let m represents the number of testing samples, among those there are n important samples. And define that:

a: the number of important samples correctly labeled as "important news story" by classifier ($0 \le a \le m$)

b: the number of outliers incorrectly labeled as "important news story" by classifier (($0 \le b \le n-m$))

Recall is defined as the number of true important samples which are determined as important ones

divided by the total number of important samples. It can be computed by following formula:

$$R = \frac{a}{n} \tag{7}$$

Precision is defined as the number of true important samples which are determined as important ones divided by the total number of samples labeled as important ones. It can be computed by following formula.

$$P = \frac{a}{a+b} \tag{8}$$

BEP value is determined by the certain point in ROC curve which precision is equal to recall. The value of BEP is set equal to the recall or precision value of that point. In our research, BEP is used to compare the performance of these three methods directly. The higher the BEP value is, the better the performance of the method is. It can be computed by the following formula:

$$BEP = P = \frac{a}{a+b}(P = R) \tag{9}$$

F1 considers both precision and recall. It can be interpreted as a weighted average of the precision and recall. In our research, F1 is used to investigate the relation ship between the number of features and performance. Generally, the higher the F1 value is, the better the performance of the method is.

$$F_1 = \frac{2PR}{P+R} \tag{10}$$

3.4. Result and Discussion

We present three contrast experiments in this part. The first experiment investigates the influence of the number of features to the performance. The second experiment is used to compare the performance of the three methods directly. At last experiment we want to find out the best k for k-means algorithm which performed the best in the second experiment.

(1) Number of features and performance

In this experiment, we set the recall in training data to 0.8 by default, and compute a threshold. In the experiment, we change the number of features from 100 to 2500 then investigate its influence on performance. F1 is used to evaluate the performance. The experiment result is shown in Figure 2.

Performance of these three methods maintained in a rational range with increasing of feature number. Thus we can see that the first 100 features are good enough to represent the whole feature space. Besides, they distributed in the "important news stories" densely. By observing these three methods respectively, k-means algorithm showed the best performance when the number of features are 300 and 400 while maintaining the performance at almost the same lever at the other numbers; Rocchio algorithm's performance keep rising with the number of features getting larger until the number of features came to 1500, after that the performance became steady; one-class SVM's performance is decreasing monotonically in the diagram with a small slope.

(2) Threshold and performance

In this experiment, we chose the number of features, 2500, 400 and 1000 which bring the best performance to relative method in the first experiment for Rocchio, k-means and one-class SVM respectively. By varying the threshold we drew the ROC curve for each method, and then BEP value is calculated. The one with highest BEP value is the best. Experiment result is shown in Figure 3.

As shown in Figure3, k-means and Rocchio are better than one-class SVM obviously. One-class SVM tried to find the hyperplane that separates important news stories dot from origin as clear as possible, therefore it is most sensitive to outliers. Because of the inevitable outliers in our training corpus, the one-class SVM showed the worst performance. In advance, k-means is better than Rocchio which verifies the assumption that the important news stories are clustered is in line with the true distribution. After

observing the news stories in training corpus, we also found that the data can easily be classified in to some clusters such as performance of company, news about important product, capital operation and so on. Therefore, it is reasonable of k-means being the best solution to our problem.

Fig. 2 Feature Number influence F1

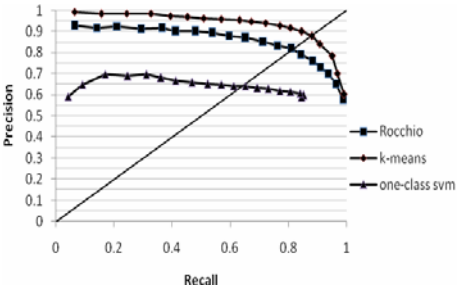

Fig. 3 ROC Curve

(3) k and performance for k-means algorithm

In this part, we did a further research on k-means algorithm which performed best on our testing corpus in order to find out the optimal parameters for it. According to the first experiment, we varied the feature number between 50 and 500 to investigate the best configuration for different k, from 1 to 9. For all the values of k, we set recall on training corpus to 0.85 (high recall means much to our research). The experiment lasted for 30 rounds and the results are shown in Table 2.

Table 2 Influence of k to Performance of k-means

k	1	2	3	4	5	6	7	8	9
F1	0.869	0.875	0.874	0.873	0.851	0.867	0.852	0.849	0.848
P	0.791	0.807	0.803	0.804	0.768	0.791	0.769	0.765	0.765
R	0.963	0.956	0.960	0.955	0.955	0.958	0.956	0.954	0.951

As shown in Table 2, when k equals to 2, 3, and 4, k-means algorithm shows the best performance which turn out precision greater than 0.8 while maintaining high recall (Because there exist outliers in training data, the threshold obtained from training data turns out smaller when it come to testing data, which results in the recall rising up by 10% in testing data). The performance can meets the needs of practical application.

4. Conclusion and Future Work

This paper used one-class classification to solve importance evaluation problem for finance news stories recommendation. Comparing to the other two one-class classification methods presented, Rocchio and one-class SVM, k-mean algorithm showed the best performance based on the training corpus generated automatically (with outliers) which guarantees a precision of higher than 0.8 while maintaining a high recall of 0.95. Besides, we proposed a new method for generating important finance news stories automatically. It can stimulate the true distribution of important finance news stories and also adapt to the change of market focus by updating the time period collected automatically.

By quantifying the importance of finance news story successfully, the method proposed here could be applied to finance search engine for ranking the results in future.

Acknowledgement

This investigation was supported in part by the National Natural Science Foundation of China (No. 60703015 and No. 60973076) and the National 863 Program of China(No. 2006AA01Z197).

References

[1] Tax, D. One-class Classification. Ph.D. thesis, Delft University of Technology, The Netherlands, 2001
[2] Scholkopf, J.C. Platt, J.Shawe-Taylor, A.J. Smola, and R.C. Williamson. Estimating the support of a high-dimensional distribution. Technical report, Microsoft Research, MSR-TR-99-87, 1999
[3] Larry Manevitz, Malik Yousef. A Novel Method for One-Class Classification Based on the Nearest Nerghbor Data Description and Structural Risk Minimization, IJCNN 2007: 1976-1981
[4] Larry Manevitz, Malik Yousef. One-class document classification via Neural Networks. Neuro Computing, 2007, 70: 1466-1481
[5] Zhang Wei, Liu Bo, Xiong Xiong. An Intrusion Detection System Basing on Outlier Mining. Systems Engineering—Theory & Pracice, 2009, 29(5): 44-50
[6] Bai Yanan, Ren Guangwei. Outlier detection in intra-day financial data case. Workshop on the Computer Network and Telecommunication. 2009: 234-240
[7] LIBSVM: http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html
[8] Koppel, M., and Shtrimberg, I., "Good News or Bad News? Let the Market Decide.", Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text, Palo Alto, CA, 2004 : 86-88
[9] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan. Language models for financial news recommendation. In Proceedings of CIKM, New York, 2000: 389–396
[10] HeXun: http://www.hexun.com/
[11] Google Finance: http://www.google.cn/finance
[12] Yahoo Finance: http://finance.cn.yahoo.com/
[13] training corpus: www.haitianyuan.com/class/news recommendation/stock/training corpus.rar
[14] testing corpus: www.haitianyuan.com/class/news recommendation/stock/testing corpus.rar