News-sensitive Multi-solution Prediction of Stock Price

Wei Fan
Nagoya University
fan@watanabe.ss.is.nagoya-u.ac.jp

Toyohide Watanabe
Nagoya University
watanabe@is.nagoya-u.ac.jp

Koichi Asakura
Daido University
asakura@daido-it.ac.jp

Abstract

Stock price analysis with data mining techniques becomes an important issue to be investigated. In this paper, we propose a system for forecasting stock prices by analyzing influence of news articles. A new kernel based possibility distribution estimation algorithm is proposed to identify multi-solution trends of stock price fluctuation. We also propose a new differentiated weighting scheme that assigns higher weights to the features if they occur in the Rise(Drop) news article cluster but do not occur in its opposite Drop(Rise) cluster.

1. Introduction

Recently, more and more research works on stock market prediction take the influence of non-quantifiable information into account. Fawcett and Provost [1] monitored the relationship between news articles and stock prices. However, a detailed procedure was absent in their paper. Other works such as [2, 3] focused on short-term fluctuation of stock price: therefore, these methods could not be applied to long-term analysis of stock price data.

In this paper, we present a system to predict the future behavior of stock market using news articles. The unique features of our system are summarized as follows: first, a new kernel based multi-solution possibility distribution estimation algorithm for trend segmentation is proposed. This algorithm satisfies flexible demand for short-term analysis of stock price fluctuation as well as long-term analysis. Second, interesting trends are clustered into two categories: rise and drop, using an hierarchical clustering algorithm based on slops and coefficient of determination. Third, a new differentiated weighting scheme that assigns higher weights to the key words (features) which occur in the Rise(Drop) news-article cluster but do not occur in its opposite Drop(Rise) cluster is proposed.

2. Stock Price Prediction System

The overview of our system is shown in Figure 1. It consists of two phases: training phase and test phase. The training phase includes six main procedures: 1) trend segmentation, which identifies trends within different time horizon; 2) trend labeling, which clusters similar trends together; 3) feature extraction, which extracts key words in news articles; 4) article-trend alignment, which associates related news articles with trends; 5) feature weighting, which assigns different weights to different features according to their performance; and 6) model generation, which generates the desired prediction model. The test phase is used to predict the future trends according to the contents of the newly broadcast news articles.

2.1. Stock Trend Segmentation

Let T be the current instant and S be the set of data points which have arrived in the time window $(T - h_t, T)$. We calculate velocity density to measure the rate of change of data concentration at a given spatial location over a user-defined time horizon h_t:

$$V_{(h_s,h_t)}(X,T) = \frac{F_{(h_s,h_t)}(X,T) - R_{(h_s,h_t)}(X,T-h_t)}{h_t}.$$

Here, $F_{(h_s,h_t)}(X,T)$ measures the density function based on the set of data points S. Similarly, $R_{(h_s,h_t)}(X,T)$ measures the density function based on the same data set reversely. Note that the velocity density is positive, if a greater number of data points which are closer to X have arrived at the end of the interval $(T - h_t, T)$. While, when a greater number of data points which are closer to X are at the beginning of the interval, then the velocity density is negative. If the trends have largely remained unchanged, then the velocity density at the location X will be almost zero.

Therefore, according to the result of velocity density, given a threshold, we can diagnose coagulation and dissolution trends at given spatial locations. Then in order to identify when these special trends occur, we compare the coagulation regions with dissolution regions at each time stamp within the time window $(T - h_t, T)$.

2.2. Stock Trend Labeling

We proposed a hierarchical algorithm to cluster the trends into different interesting categories based on: 1) the slope of the segment (m) and 2) the coefficient of determination (R^2). Each segment is thus represented by (m, R^2). The segments are merged according to the minimum group average distance $GAD(C_i, C_j) = \frac{\sum_{i \in C_i} \sum_{j \in C_j} d_{ij}(i,j)}{|C_i||C_j|}$, where $|C_i|$ and $|C_j|$ are the magnitudes of the clusters C_i and C_j respectively; $d(i,j) = \sqrt{(m_i - m_j)^2 + (R_i^2 - R_j^2)^2}$ is the Euclidean distance between the objects inside C_i and C_j. The clustering procedure terminates when the number of clusters are equal to three (Rise/Drop/Steady). Those segments in the cluster having the maximum average slope are labeled as Rise. Similarly, those segments in the cluster having the minimum average slope are labeled as Drop. Segments in the remained cluster are labeled as Steady.

2.3. Article and Trend Alignment

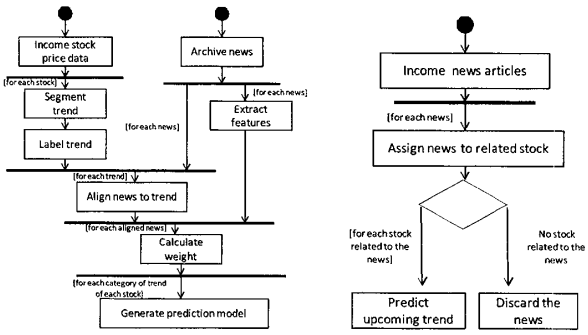
Let T-clusters be the clusters on trends and N-clusters be the clusters of news articles. First, all of the news articles that are broadcast within T-cluster Rise (Drop) are grouped together. Each article is represented by a normalized vector-space model $d_i = (w_1, w_2, \cdots, w_n)$, where the element w_t corresponds to the score of key word t in the article d_i, and it is calculated by the standard $tf \cdot idf$ scheme: $w_t = tf_{d,t} \times \log \frac{N}{df_t}$, where $tf_{d,t}$ is the frequency of t in the article d; df_t is the number of articles containing the term t; N is the total number of articles contained in the particular T-cluster (Rise/Drop). Incremental K-Means is then used for splitting the weighted article into two clusters. The centroid of the cluster C_i is defined as $C_i = \frac{1}{|S_i|} \sum_{d \in S_i} d$, where S_i is the set of articles within the cluster C_i and $|S_i|$ is the number of articles in this set. The similarity between the article d_i and the centroid C_j is determined by the cosine measure: $cos(d_i, C_j) = \frac{d_i \cdot C_j}{|d_i||C_j|}$, where $|d_i|$ and $|C_j|$ is the magnitude of the article d_i and the cluster C_j respectively.

2.4. Differentiated News Article Wighting

In order to differentiate the features appearing in one of the cluster but not the other, two coefficients are introduced: inter-cluster discrimination coefficient: $CDC = \frac{n_{i,t}}{N_t}^2$ and intra-cluster similarity coefficient: $CSC = \sqrt{\frac{n_{i,t}}{n_i}}$, where $n_{i,t}$ is the number of articles in the N-cluster i containing key word t; N_t is the number of articles containing t; and n_i is the number of different key words.

2.5. Learning and Prediction

The association between different features and different category of trend is generated based on Support Vector Machine (SVM). We have a pair of classifiers. One is responsible for classifying whether a news article will trigger a rise event, the other is responsible for the event of drop.



(a) Training phase (b) Test phase

Figure 1. The system overview

3. Conclusion

In this paper we demonstrated a system which monitors the stock market and predicts its future behaviors. A velocity density estimation algorithm and a hierarchical clustering algorithm are introduced. An incremental K-Means algorithm is used to identify the important features within the collection of articles. In our future work, we will do experiments to evaluate the performance of our approach.

References

[1] T. Fawcett and F. Provost: Activity Monitoring: Noticing Interesting Changes in Behavior. In Proceedings of the 5th International Conference on KDD, San Diego, (1999)

[2] M.A. Mittermayer and G. Knolmayer: Text Mining Systems for Market Response to News. A Survey, Working paper (2006)

[3] Seo, Y.W., Giampapa, J.A. and Sycara, K.: Financial News Analysis for Intelligent Portfolio Management. Technical Report CMU-RI-TR-04-04, Carnegie Mellon University, (2004)