Building An Expert System Based On Data Mining

Sagar Bhushan Gawde,
Department Of Computer,
Vidyalankar Institute Of Technology,
Wadala East- 400037, Mumbai, India.

Abstract— This work presents a novel framework for predicting stock trends and making financial trading decisions based on a combination of Data and Text Mining techniques. The prediction models of the proposed system are based on the textual content of time-stamped web documents in addition to traditional numerical time series data, which is also available from the Web. The financial trading system based on the model predictions uses three different trading strategies. In this work, our system is simulated and evaluated on real-world series of news stories and stocks data using Decision Tree Induction Algorithm. The main performance measures are the predictive accuracy of the induced models and, more importantly, the profitability of each trading strategy using these predictions.

Introduction

The Efficient Market Hypothesis (EMH), as stated by Fama ([7], [10], [11]), assumes that 'Stock prices fully reflect all their relevant information at any given point in time'. As the basis for growth and development of a modern economy this means that no information or analysis can be expected to outperform the market and that stock prices follow 'Random Walks' ([9]), where a change in stock price over time is purely random and statistically independent of the stock price in the past. However, to this day no one can explain the anomalies in the market, which can be utilized to assure some short term predictive power ([6], [9], [12]). In making their own forecasts most financial specialists try to exploit the time gap of the market’s adjustment to new information. They reduce their risk by combining both technical (base future price predictions on past prices) and fundamental (base predictions on real economy factors, such as inflation, trading volume, organizational changes in the company etc.) analysis strategies, which are mentioned by Gidofalvi ([4]) and fully explained by [9]. In making their own forecasts most financial specialists try to exploit the time gap of the market’s adjustment to new information. They reduce their risk by combining both technical (base future price predictions on past prices) and fundamental (base predictions on real economy factors, such as inflation, trading volume, organizational changes in the company etc.) analysis strategies, which are mentioned by Gidofalvi ([4]) and fully explained by [9]. In order to obtain the data required by both strategies, one can refer to various publicly available resources like the stock market itself, the companies, news articles, etc. A rather new source for information in the late 20th and the 21st centuries is, of course, the Internet. In order to exploit this relatively new media as an additional tool supporting the forecasting task, we need to combine techniques from both time series data mining and web content mining.

In this work, we present a new system for detecting stock trends based on the combination of Data Mining and Web Content Mining techniques. It is a new Financial Trading System which:

1) Creates a “melting pot” of numeric and textual data before running an induction algorithm,
2) Extracts automatically key words and phrases instead of using a prior expert list of phrases,
3) Eliminates the need for word independence assumption by using Decision Trees rather than Naive Bayes,
4) Extends the Window of Influence of news articles in the prediction task to days

Equations

\[ S = \frac{1}{3} \cdot \frac{TF}{N} + \frac{1}{3} \cdot \frac{P}{L} \cdot \frac{B}{L} + \frac{1}{3} \cdot AV \ldots \] (Eq.1)

Where:
- L: the time frame, in days, for the word dictionary.
- B: the time window between the first and last occurrence of a word.
- P: the number of days to the last occurrence of a word.
- TF: the number of occurrences of a word during L (known as Term Frequency).
- N: the number of words in the dictionary.
- AV: the annualized volatility of the stock as calculated by

\[ \sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (u_i - \bar{u})^2} \] (Eq.2)

Where \( u_i \) is the \( i^{th} \) observation of the stock price and \( \bar{u} \) is the mean of stock price.

Whereas the proposed system uses

\[ S = \frac{1}{3} \cdot \frac{TF}{N} + \frac{1}{3} \cdot \frac{P}{L} \cdot \frac{B}{L} + \frac{1}{3} \cdot AV \ldots \] (Eq.3)
Proposed Model

Our system is designed as a full cycle prediction system for stock trends according to past numeric values of the stocks as well as their related textual web articles. It goes through six steps, as shown in Fig. 1, which are:

- **Step 1**: Data Collection from the Web.
- **Step 2**: Feature Extraction.
- **Step 3**: Textual Weighting.
- **Step 4**: Combined Data-Set Construction.
- **Step 5**: Classification Model (Decision Tree) Induction.
- **Step 6**: Market Action Recommendation.

### Experimental Results

From the experiments carried out on web. Words are collected for a selected stock for a specified period, this step includes Textual extraction from authentic sites. The chart of the selected stock can also be analysed in this step.

#### Conclusion

This work demonstrates a new paradigm for the advance reading of stock movements and helping financial decisions in trading resting on a mix of financial information and news mining methods. The trend deciding factors of the proposed system make use of text material that is dated present in web pages along with customary numeric price time series information, present on the Internet. The proposed trend deciding system makes use of more than one...
system thereby increasing chances of accurate decisions. The method is simulated and assessed based on actual time series data and news information.

References


[7]. E.F. Fama, Long Term Returns and Behavioral Finance, Social Science Research Network.


