Data Mining Project Proposal

Konstantin Kutzkov (konk@itu.dk)
Kevin Tierney (kevt@itu.dk)

March 31, 2011

Project Question

Does past stock performance really matter?

Data Sources

We will use publicly available data in Yahoo Finance to get historical stock prices and volumes on a day to day basis for the past several years. We can combine this data with data from ScotTrade, which is behind a paywall, but for which Kevin has an account. We can calculate the technical indicators we wish to investigate based on the stock price and volume.

Data Mining

Supervised Learning

Using a feature vector of technical indicators and prices we will learn when to buy and sell a stock. The labels for data points will be either "buy", "sell", or "hold" which we can determine by reading the stock prices and determining which periods of time would make money and which would lose money if we held the stock. We may try to have "strong buy" and "strong sell" as labels in order to have more human useful information come out of the learning.

We plan to try support vector machines along with neural networks for this stage of the research. We will experiment with different kernels for the SVM and different hidden layer sizes for the neural network.

Frequent Pattern Mining

We propose to attempt to determine rules about stock movements by extending the approach presented in [2]. In [2] each day is assigned a category based on the price movement of the stock. These categories can then be mined to determine rules. We will experiment with adding more categories based and more features of a stock price behavior, e.g. volume of sells, stochastic oscillator for a given period of time.

We plan to use a publicly available state-of-the sequential mining solver, e.g. Prefix Span [1]. However, the exact representation as event sequences is yet to be decided.

Time Plan

- Mid April: Prepare full dataset
- End April: Supervised Learning results
- Early May: Frequent pattern mining results
- May 18: Hand in report

References

[1] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Jianyong Wang, Helen Pinto, Qiming Chen, Umeshwar Dayal, Meichun Hsu. Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach. *IEEE Trans. Knowl. Data Eng. 16(11)*: 1424–1440 (2004)

[2] Jo Ting, Tak-Chung Fu, Fu-Lai Chung: Mining of Stock Data: Intra- and Inter-Stock Pattern Associative Classification. *DMIN 2006:* 30–36