40th Southeastern Symposium on System Theory
University of New Orleans
New Orleans, LA, USA, March 16-18, 2008

Supervised and Unsupervised Methods for Stock Trend Forecasting

Nicole Powell, Simon Y. Foo, and Mark Weatherspoon

Department of Electrical and Computer Engineering
FAMU-FSU College of Engineering
Tallahassee, Florida 32310, U.S.A.
{powelni4, foo, weathers}@eng.fsu.edu

Abstract – Stock forecasting is a major component of any finance institution because predictions of future prices, indices, volumes and many more values are often incorporated into the economic decision-making process. Although there are many different approaches out there, this paper will compare unsupervised classification techniques such as k-means clustering with supervised learning algorithms such as support vector machines (SVMs). In our study, a list of stock prices taken from historical data of the S&P 500 is used as our testbed. These prices will be categorized as increasing or decreasing in price on a weekly basis. The goal of this study is to determine the best method for forecasting the trend of stock prices.

Keywords: Pattern recognition, stock market prediction, time series, supervised learning, unsupervised classification, k-means clustering, support vector machines.

I. INTRODUCTION

The prediction of any aspect of the future has always fascinated mankind because of the possible benefits of this knowledge, especially when dealing with finances. For recognizing specified patterns, is it important to develop a method for eliminating speculation and to investigate new algorithms for detecting patterns. There has been much work done in this area however no real successful formula has been developed [1]. From year to year, many stockholders would like to be able to know if their stock price will increase or decrease, and in turn this prediction may help them in the decision to buy more or sell the stock they currently have. It is widely known that the stock market is a volatile and complex entity which is affected by various factors such as government policies, political situations, public events, internal company politics and much more. However, we have no way of knowing exactly which factor will affect stock prices nor do we know how it will affect the trend of that stock. In the task of stock trend prediction it is more natural and effective to represent the target values by the successive relative changes in price since the previous time point[2].
In this paper the stock trend prediction is approached as a technical analysis problem using K-means clustering and support vector machines for classification. Technical analysis ignores any underlying factors, such as company profit or market sector, and focuses on finding patterns directly from the stock data. The results from the two methods will be compared to determine which method, if either, can accurately predict the trend of stock prices.

A. Data Collection and Problem Formulation

The data used in the analyses contained the weekly prices of Dow 30 stocks. This data was created from historical data obtained from Yahoo! Finance [3]. We choose to use two years of data to identify the performance. In particular the first year, 2005, will be used for training and the second year, 2006, for validation.

There are nine attributes associated to each stock including the decision on whether the stock's price increased or decreased during a 52-week period. The attributes are raw numerical price values as follows: open, high, low, close, volume, volume increase or decrease, adjusted close, adjusted close increase or decrease, and overall increase (Class +1) or decrease (Class -1) over the 52 weeks.

This problem is then formulated as a two-class pattern recognition task. The classification algorithms needed to simply determine if the stock price would increase or decrease based on the given data from the previous year.

II. LITERATURE REVIEW

As stated earlier, in general the approaches to predicting the stock market can be classified into two classes, fundamental analysis and technical analysis. Various solutions to stock market prediction problems have been designed in the past, including support vector machines. Most of these published works are targeted at overseas markets, most commonly the Australian Stock Exchange. Very little is known or done on predicting the US stock market in the area of pattern recognition.
Unsupervised pattern recognition, particularly using K-means clustering, has not been fully explored. This is because it is difficult to predict behavior patterns from old data. When data is time variant, data selection will influence the training result [4]. Some unsupervised

methods that have been explored include radial basis functions and self-organizing maps. iJade Stock Advisor is a stock prediction system that was developed using a RBF recurrent network. It is discussed in more detail in [5].

Some supervised methods that have been explored include amnestic neural networks and linear SVMs. Traditional neural networks do not solve this problem so amnestic neural networks were designed and described in detail in [4]. Amnestic neural networks are an extension of backpropagation; generally speaking the effectiveness of this model depends on present data being more useful than data from long ago. (Old data has less effect on the training result, just like gradually forgetting.)

SVMs address the problem that minimization of empirical error does not guarantee small actual error; empirical risk minimization is described in detail in [6]. Using the Australian Stock Exchange, SVMs were used strictly for stock selection. This was an attempt to identify stocks that are likely to outperform the market by having exception returns. The equally weighted portfolio formed by the stocks selected by SVMs had a total return of 208% over a five year period, significantly outperforming the benchmark of 71% [6].

III. METHODOLOGY

Two classification techniques are described in detail below. The data is separated into a training set which consists of 2005 Dow 30 stocks and a testing set which consist of 2006 Dow 30 stocks.

A. K-means Clustering

The K-means clustering algorithm is an unsupervised learning method and is discussed in detail in [7]. The K-means clustering algorithm was written in MATLAB (found in Appendix A). First the data is normalized to prevent the weights of different attributes of the data from affecting the results. The formula for normalization is:

$$\frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Principal component analysis (PCA) is used to reduce the number of dimensions of a data set to 'select' the components which have the biggest effect when classifying the data set. More detail on PCA is described in [8]. The basic steps for K-means clustering are as follows:

i. Initialize the centroids, which shold be equal to the number of classes (targets) in the data set.
ii. Determine the distance from each object to each of the centroids.
iii. Select the minimum distance and group each object based on the minimum distance.
iv. Continue the previous steps until none of the data points are reassigned.

The distance metrics used were Euclidean, Manhattan, Chebyshev, Minkowski and Canberra. The results involve analysis of which distance metric provided the best results/classification.

B. Support Vector Machines

Support vector machines (SVMs) are a supervised pattern classification technique which map input vectors to a higher dimensional space where a hyperplane is constructed. SVMs can be used on linearly and non-linearly separable data. On a straight line, a hyperplane is a point which divides a line into two rays. In 2D, a hyperplane is a line which divides the plane into two half-planes. In 3D, a hyperplane is an actual plane which divides the space into two half-spaces. SVMs construct two parallel hyperplanes on each side of the hyperplane that separates the data. The separating hyperplane increases the distance between the original hyperplanes. The greater the distance between the original hyperplanes will result in a smaller amount of error in classification. More on SVMs can be found in [6]. The software used in this analysis can be found at [9].

IV. EXPERIMENTAL RESULTS

A. K-means Clustering

K-means clustering was applied to the data using all five distance metrics. Figure 1 shows the relationship between the number of components in relation to the accuracy of the pattern recognition system for each of the distance metrics. Table 1 shows the average accuracy for each distance metric.

TABLE 1
K-MEANS CLUSTERING ACCURACY AVERAGES

Distance Metric	Accuracy
Canberra	79.8%
Minkowski	74%
Chebyshev	72.1%
Manhattan	71.3%
Euclidean	71%

The Canberra distance metric performed the best with an average accuracy of 79.3% and the highest accuracy value of 85.9%. This maximum accuracy value obtained used only one component. The Euclidean metric proved to be the worst metric with an average accuracy of 71%. However, the Manhattan metric encompassed the lowest accuracy value of 62%.

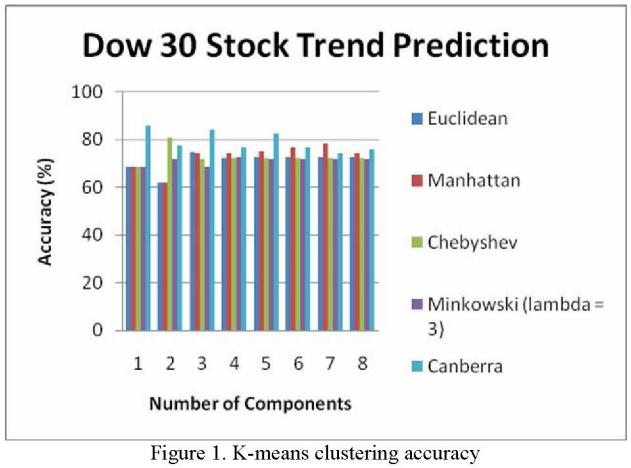

Figure 1. K-means clustering accuracy

B. Support Vector Machine

The highest accuracy value obtained using the support vector machine program was 89.1%. The accuracy of the algorithm also depends on which kind of kernel function. Table 2 shows the results.

TABLE 2
ACCURACY VALUES FOR SVM SOFTWARE

Kernel Method	Accuracy
Linear	89.1%
Gaussian	84.7%
Polynomial	76.5%

V. CONCLUSION

Both K-means clustering and support vector machines proved to be the comparable methods for this application. The best accuracy value was 89.1% for SVMs and 85.6% for K-means clustering; which are both fairly close values.

Therefore it can be concluded that an unsupervised method can be used for stock trend forecasting because unsupervised methods find patterns in data that is usually seen as uncorrelated. Uncorrelated definitely can be used to describe the stock market because there are many factors that we assume affect stock prices but we do not know for sure.

It also seems that linear support vector machines prove to be very useful. Although the data is not easily separable nor is it related in any way, SVMs were still able to applied.. Typically a linear kernel SVM will result in higher results than that of K-means clustering in other applications, when dealing with a two-class database. That may not be the case here because the data may not be enough to classify stock trend forecasting. Also, there were not equal amounts of the classes which may alter the results. In order to better classify the database overall, the data should be more evenly distributed.

VI. ACKNOWLEDGEMENTS

A special thanks to Dr. Simon Foo and Dr. Mark Weatherspoon for providing invaluable guidance.

REFERENCES

[1] Lee, Jae Won. Stock Price Prediction Using Reinforcement Learning. ISIE 2001.

[2] Afolabi, Mark O and Olatoyosi Olude. Predicting Stock Prices Using a Hybrid Kohenen Self Organizing Map (SOM). Proceedings of the 40th Hawaii International Conference on System Scienes, 2007.

[3] Yahoo! Finance. www.finance.yahoo.com

[4] Ye, Qiang, Bing Liang, and Yijun Li. Amnestic Neural Network for Classification: Application on Stock Trend Prediction. IEEE 2005.

[5] Lee, Raymond S. T. iJADE Stock Advisor: An Intelligent Agent Based Stock Prediction System Using Hybrid RBF Recurrent Network. IEEE Transactions on Systems, Man, and Cybernetics. Part A – Systems and Humans, Vol. 34, No. 3, May 2004.

[6] Fan, Alan and Marimuthu Palaniswami. Stock Selection using Support Vector Machines. IEEE 2001.

[7] Faber, Vance. Clustering and the Continuous k-Means Algorithm. Los Alamos Science. Vol 22. 1994. pp 138-144. http://www.fas.org/sgp/othergov/doe/lanl/pubs/00412967.pdf

[8] Shlens, Jonathan. A Tutorial on Principal Component Analysis. Systems Neurobiology Laboratory, Salk Insitute for Biological Studies La Jolla, CA 92037 and Institute for Nonlinear Science, University of California, San Diego La Jolla, CA 92093-0402. Dated: December 10, 2005; Version 2. http://www.cs.cmu.edu/~elaw/papers/pca.pdf

[9] Anguita, D.; Ridella, S.; Sterpi, D. A new method for multiclass support vector machines. Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on Volume 1, Issue , 25-29 July 2004 Page(s): - 412