A SVM APPROACH TO STOCK TRADING

Christopher King   Christophe Vandrot   John Weng

INTRODUCTION

Statistical arbitrage is the application of modern statistics, significant computing power, and large data sets to the discovery of financial market mispricing and then exploiting those inefficiencies for profit. Fundamental analysis is the process of analyzing all public company disclosures including financial statements in the attempt to uncover indicators of future company performance. In this paper, we endeavor to combine both approaches to investing in pursuit of superior risk adjusted returns.

There are a large number of market participants with different views on the correct valuation of varying securities and these participants constantly update their own market outlook and security views. The information a market participant may consider relevant on one day may be considered insignificant by the same individual one month later. The markets are in a constant state of flux of what is critical for security valuation and the monetary value to ascribe to this information.

Support Vector Machine (SVM) is currently one of the most popular learning algorithms and has successfully been applied to numerous fields. Its popularity can be explained both by a high generalization performance and a mathematically well posed training method. In the context of financial data modeling, SVM is particularly appealing for multiple reasons. Firstly, it makes no strong assumptions on the data. Secondly, since it is not an empirical error minimization method, therefore it should not over-fit the data. These two reasons explain why SVM compares well to other methods commonly used for financial forecasting such as ARIMA of Artificial Neural Networks (See [CaoTay01]). Further, in our context of incorporating several features, we need an algorithm that handles high dimensional data well, which is the case for SVM.

OBJECTIVES

We chose to focus on the oil sector of the S&P500 stock index. There are seven stocks in this group including well known firms such as Exxon Mobil and Hess Corp. This group of companies offers homogeneous products leaving them to compete on superior business execution and overall firm financial management. Since there is no significant product differentiation, we believe security valuation is highly related to financial statement metrics indicating company execution that investors continually decide are more and less important during the evolution of time.

We chose to use the SVM to classify the relative performance of the stock against the performance of other stocks in the sector. Then we manage a portfolio making investments decision entirely based on the classification made by the learning algorithm. We long (buy) the stocks that are expected to outperform, and short (sell) the stocks that are expected to underperform the sector. Using this approach, we are not buying stocks on the expectation that they will rise and selling stocks on the expectation the stock price will fall. Instead, we are buying stocks on the expectation that the price of the outperformers will rise more than the price of the underperformers or the price of the outperformers will fall less than the price of the

underperformers. In this manner, we are only predicting the relative performance of one portfolio, outperformers, relative to the other, underperformers.

The following table includes the fundamental measures we used to input into the SVM algorithm. We surveyed several academic papers ([AbarbanellBushee1998], [AbarbanellBushee1997], [LevThiagarajan1993], [Nicoletti2004]) that described the connection between several of these measures and stock prices. Furthermore, we included obvious measures like historical oil price, which should have a strong correlation to the stock prices for companies in the oil sector, or information about the trading volume. Also, the features are presented in what we expected to be the decreasing order of importance of the features.

Inputs	Underlying Connection
Historical Price	
Trading Volume	
Historical Oil Prices	Large driver of industry costs
P/E Ratio	Number of years of earnings equivalent to the current company value
Enterprise Value / EBITDA	Metric widely used in private equity to gauge company value
Current Assets / Current Liabilities	Indicates whether the company may have near term financial distress
Total Assets / Total Liabilities	Rough indication of the residual value of the company
Percentage of Analysts Recommending Buy and Sell	Reflects the opinions of company outsiders that closely follow the firm
Investing Cash Flow / Depreciation	Indicates whether the firm is adequately investing for the future
Market Capitalization / Operating Cash Flow	Similar to PE ratio, but with an accounting metric not easily manipulated
Market Capitalization / Revenue	The value of the firm relative to $1 of revenue
Operating Cash Flow / Revenue	Percentage of each revenue dollar the firm turns into operating cash flow
Net Increase (Decrease) Cash / Revenue	Percentage of each revenue dollar the firm turns into cash

DESIGNING OUR EXPERIMENTS

An SVM classifier was applied on each one of the 7 stocks in the oil sector. Our data set contains daily values of the inputs presented in the table above during the period from January 2001 to November 2009. Our goal was to forecast what would happen on a particular day using only measures obtained from the 5 previous days. We used the first 6 years of data to train the algorithm then made forecasts and invested on a daily basis during the remaining 3 years.

Using twice the amount of data for training than for testing is a common practice when applying SVM classifier. Further, we believe that the previous 5 days convey enough information to forecast the next day's performance.

The inputs we selected were transformed whenever we believe the evolution conveyed more information than the absolute value itself. To us, past stock prices, past index, and trading volume all fall in that category. The first two were transformed into daily returns and volume was used to derive the difference of the current volume and a moving average mean.

The SVM algorithm was specified to use the Gaussian kernel. The resulting algorithm therefore has two parameters: the cost of misclassifying and the bandwidth of the kernel. These two parameters were determined using a 10-fold cross validation method.

The results we present in the next paragraph were obtained using the LS-SVMlab Toolbox for Matlab referenced [ToolBox][LSSVM].

RESULTS

Here we present the results we obtained using respectively 1, 5 and 14 features.
The forecasting performance is presented in the table below.

The algorithm generally predicts correctly more often than incorrectly. Also, having more features tends to give a better rate of accurate forecasts but the difference of performance is low.

Stock		1	2	3	4	5	6	7
Number of features								
Percentage of accurate forecasts	1	51.66 %	50.07 %	51.93%	49.67 %	51.93 %	51.53 %	50.33%
	5	54.59%	50.33 %	52.20 %	51.26 %	48.87%	51.00%	51.00%
	14	54.06 %	51.40%	50.87 %	51.80 %	49.67%	52.46 %	52.60%

Having a better rate with more than one feature tells us that the additional features would contain more information than the stock price itself.
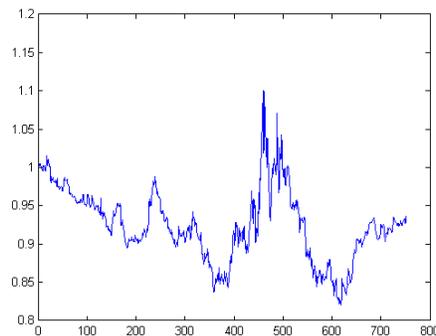Our conclusion is that adding the last 9 features is not worth the additional complexity.

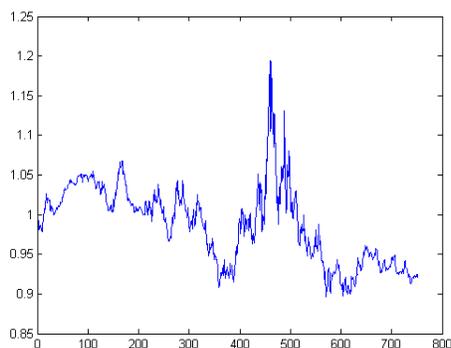Next, we present the evolution of the corresponding portfolios.

Using 1 feature:
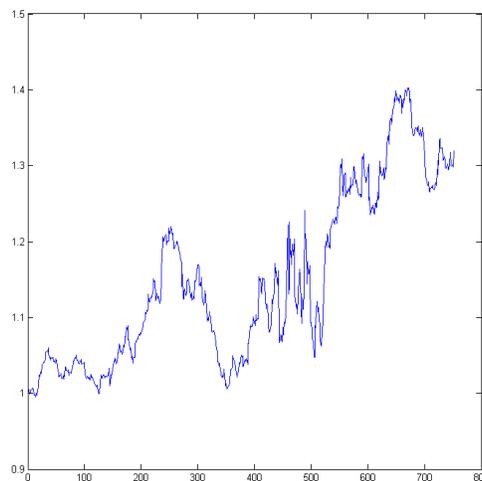
Using 5 features:

Using 14 features:

The performance of the first portfolio is misleading. One should note that the algorithm did not perform well in terms of forecasting with only one feature. Also, looking at the results using 5 and 14 features, it is hard to see whether the trading strategy would be appropriate only up to some limited time after the last date of the training set.

CONCLUSION

An SVM classifier was successfully applied to forecast relative performance within the oil sector. The forecasts made are good in terms of the percentage of accurate guesses. Also, our results tend to show that the features we selected do contain information relating to future performance. Furthermore, it was clearly illustrated that those predictions, even though they have an acceptable rate of good "guesses" would not necessarily lead to profitable strategies.

There are many reasons why we believe that the application of SVM to financial forecasting is still very promising. Further work should focus on investigating the selection of other meaningful inputs among which could be price volatility, current trend, and other financial ratios or out of the sector market factors. It would be beneficial if additional transformations such as returns over several periods, scaling, detecting and removing outliers from the data set be carried out using the inputs. See [CaoTay2001] for more details.

In order to derive a better trading strategy, one could consider applying this method over longer time scales. As a matter of fact, we placed ourselves in the position of over-trading: our

portfolio's rate of return would have suffered a lot from taking in to account transaction costs. On the other hand, forecasting further in the future is more difficult because one no longer benefits from market persistence.

Additional improvements could also address how to adapt SVM to the specific purpose of forecasting financial data. Researchers L. Cao and F. Tay have devoted many research papers to this topic, suggesting innovative ideas such as giving more weight to recent training data [TayCao2002a], or experts [Cao2003] to take non-stationarity into account.

REFERENCES

[AbarbanellBushee1998] Abnormal returns to a Fundamental Strategy, Abarbanell, Bushee, The

Accounting Review, 1998

[AbarbanellBushee1997] Fundamental Analysis, Future Earnings and Stock Prices, Journal of

Accounting Research, 1997

[LevThiagarajan1993] Fundamental Information Analysis, Journal of Accounting Research,

1993

[Nicoletti2004] On the Profitability of 12 Fundamental Analysis Strategies on the NASDAQ 100,

Nicoletti, 2004

[CaoTay2001] Financial Forecasting Using Support Vector Machines,

[TayCao2002a] ϵ-Descending Support Vector Machines for Financial Times Series Forecasting,

Tay, Cao, Neural Processing Letters, 2002

[Cao2003] Support Vector Machines Experts for Times Series Forecasting, Cao,

Neurocomputing, 2003

[ToolBox] LS-SVMlab, Pelckmans et al., http://www.esat.kuleuven.be/sista/lssvmlab/

[LSSVM] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle,

Least Squares Support Vector Machines,

World Scientific, Singapore, 2002 (ISBN 981-238-151-1)