CAN MACHINE LEARNING TECHNIQUES BE USED TO PREDICT MARKET DIRECTION? - THE 1,000,000 MODEL TEST

Jonathan Kinlay, PhD
Systematic Strategies LLC
590 Madison Avenue, 21st Floor
New York, NY 10022
tel: +1 (212) 521 4303
email: jkinlay@systematic-strategies.com

Dan Rico, MSc, FnEng
tel: +1 (905) 761-9366
email: dan_rico_to@yahoo.com

1. INTRODUCTION

Over the years several studies have reported what appear to be promising results produced by nonlinear classification and machine learning techniques such as Support Vector Machines (SVM), Random Forests, and Neural Networks which can be used for market prediction. Examples include Kim (2003), Cao and Tay (2003), Huang, Nakamori and Wang (2004), and Wang and Zhu (2008). Model inputs typically comprise price and volume data, and may also include a selection of well-known technical indicators, which some studies have been reported to be influential – see, for example, Rao and Hong (2010).

By and large such studies have tended to focus on developing models to forecast market direction rather than returns, while other research has indicated that, in principle, the ability to forecasting the sign of future returns should alone be sufficient to support profitable trading strategies. Levich (2001), for example, focuses on what he terms *useful* forecasts, which predict the direction of price change and hence "... lead to profitable speculative positions and correct hedging decisions. "There is ample evidence that sign forecasting can accomplished successfully. Relevant literature includes Breen, Cheung, Chinn and Pascual (2003),Elliott and Ito (1999), Gencay (1998), Glosten and Jagannathan (1989), Kuan and Liu (1995), Leitch and Tanner (1991),Leung, Daouk and Chen (1999), Pesaran and Timmermann (1995), Pesaran and Timmermann (2001), Shellans and Paul (1992), Wagner, Larsen andWozniak (1995), Womack (1996), and White (2000).

Critics have raised doubts about these techniques on several grounds including, in the case of neural networks, the difficulty in articulating the model structure and function in precise mathematical terms. Another criticism is the potential for bias in the sample selection: because of the complexity and intensity of the computation effort involved, out-of-sample testing is sometimes conducted on small data sets that are unlikely to be representative of the full range of market behaviors. A further criticism is that the evaluation has in some cases been limited only to the prediction accuracy rate of the classification engine and often does not include performance measurements of an associated market timing trading strategy. As result, the performance scores published in the literature can be misleading and do not accurately mirror the potential losses that the trading strategy may incur. We might characterize some of the criticism as being targeted more broadly at "data mining" techniques in general, on the grounds that, over time, if enough models are thrown at enough data some positive results are likely to turn up eventually simply by random chance.

Anticipating that a more comprehensive study might produce a definitive answer to the question of whether the machine learning approach really does have predicative power, in this paper we conduct an analysis of the ability of a wide variety of nonlinear models to forecast market direction in the S&P 500 Index and selection of liquid component stocks.

We were assisted in the research by the 11Ants and Matlab software programs, which enable large-scale simultaneous model construction and testing using a wide variety of modeling techniques, including all those mentioned above. The software tools enable the researcher to examine the performance of individual models and ensembles of models combining several different forecasting techniques.

2. RESEARCH PROGRAM AND DATA

Our research was carried out in two distinct programs. In the first we focused exclusively on applying SVM techniques to forecast market direction in the S&P 500 index (^GSPC) and a total of ten blue chip stocks from different sectors over the period from January 2000 to November 2010. The stocks selected for testing were:
Microsoft (MSFT), Google (GOOG), Apple (AAPL), IBM, Citigroup (C), Bank of America (BAC), Goldman Sachs (GS), Potash (POT), Mosaic (MOS) and Goldcorp (GG).

In the second research program we focused exclusively on the SPY ETF, using a competitive model framework provided by the 11Ants modeling system to select the best performing combinations of non-linear models employing a variety of non-linear classification techniques, applied to a data set comprising daily high, low, open and close prices and trading volume, over the period from November 1993 to January 2011.

In addition to price and volume data, and prior-period returns, a comprehensive of well known technical indicators and oscillators were incorporated into universe of potential model inputs, including [3]:

1. EWMA (7, 50 and 200 days)
2. Moving Average Convergence/Divergence (MACD)
3. Relative Strength Index (RSI)
4. Average Directional Index (ADX)
5. Fast and slow stochastic oscillators
6. Momentum and acceleration
7. Williams R%
8. Accumulation/distribution oscillator
9. Chaikin oscillator
10. Williams accumulation/distribution

3. RESULTS

PART 1
The goal of our research is twofold:

1. validate previous published SVM performance results and perform additional back-tests and performance scores; and
2. identify an optimal set of technical indicators with the largest predictive power for individual securities from specific sectors.

In order to evaluate the prediction power of machine learning algorithms we selected the SVM approach as our principal analytical tool. The goal of SVM classification engine is to provide an optimal method of separating hyperplanes in high dimensional feature space.

The success of the SVM in various engineering fields is due, but not limited, to the following advantages:
- reduces the risk of overfitting: small number of parameters to tune (C and sigma) compared to the back propagation neural network
- less sensitivity to noise
- dimensionality is not an issue

On the other hand some of the SVM drawbacks are:
- the process of kernel function selection can be cumbersome
- require a lot of memory and CPU time

SVM performance was tested for the S&P 500 index and the selected blue chip stocks using as features the normalized low and high, adjusted close prices and the above first eight technical indicators.
The following parameters were used for the technical indicators:
- RSI: number of periods – 14
- ADX: samples lagging average - 7, 50, 200
- fast stochastic oscillators: number of periods – 10
- slow stochastic oscillators: number of periods – 3
- momentum: number of periods – 4
- Williams %R: number of periods – 14

The SVM tests were carried out using a Gaussian Radial Basis Function as kernel function holding two possible sets of optimal values for the parameters C and sigma while the input data was normalized to zero mean and scaled to unit variance:
a. C = 78, sigma = 0.1414
b. C = 100, sigma = 0.071

In order to cover a large variety of possible backtesting scenarios, the input of 2713 data entries was split into training and testing data sets as follows:

1. fixed number of training and testing data sets

   a) 2683 training and 30 test observations
   b) 2533 training and 180 test observations

2. rolling window

   a) training sample of 1000 entries and 30 testing days
   b) training sample of 1000 entries and 30 testing days and add the previous testing data to the learning pool

   c)

To identify an optimal set of technical indicators with the largest predictive power, the tests were performed using all possible features combinations. For each individual security the combination of technical indicator features that produced the highest SVM accuracy prediction score was selected. The SVM performance results for different parameter settings, testing data and optimal technical indicators are shown in Tables 1-4 below.

SVM Parameters: C = 78, sigma = 0.1414		
Test data: last 30 days		
Security	Accuracy Prediction	Feature Index
^GSPC	0.7	1,3,4
MSFT	0.77	1,2,3,8
GOOG	0.77	1,5
AAPL	0.6	1,2,3,4
IBM	0.77	4,6
C	0.63	1,2,4,6
BAC	0.67	1,2,5,6,8
GS	0.73	3,6,8
POT	0.7	1,5,6
MOS	0.7	1,2
GG	0.7	1,2,6

Table 1

SVM Parameters: C = 78, sigma = 0.1414		
Test data: last 180 days		
Security	Accuracy Prediction	Feature Index
^GSPC	0.59	1,3,7,8
MSFT	0.62	5,8
GOOG	0.59	2,6
AAPL	0.59	3,5
IBM	0.61	3,6,7
C	0.56	1,2,6
BAC	0.56	2,5,6,8
GS	0.59	2,3,8
POT	0.53	5,7
MOS	0.61	1
GG	0.54	2,8

Table 2

SVM Parameters: C = 100, sigma = 0.071			
Test data: last 30 days			
Security	Accuracy Prediction	Feature Index	Sharpe Ratio
^GSPC	0.73	2,6,7,8	6.37
MSFT	0.8	1,3,7	6.64
GOOG	0.7	6,8	4.82
AAPL	0.6	1,2,3,4	3.08
IBM	0.77	1,2,3,4	5.09
C	0.7	5,8	4.14
BAC	0.63	1,5,6,7	1.92
GS	0.77	7,8	9.76
POT	0.63	1,5	4.16
MOS	0.67	1,2	6.88
GG	0.63	2,5	3.76

Table 3

SVM Parameters: C = 100, sigma = 0.071		
Test data: last 180 days		
Security	Accuracy Prediction	Feature Index
^GSPC	0.59	1,2,8
MSFT	0.62	6,7,8
GOOG	0.58	1,7
AAPL	0.59	1,2,4,5
IBM	0.62	3
C	0.58	7,8
BAC	0.58	3
GS	0.53	6
POT	0.53	6,8
MOS	0.59	1
GG	0.56	5,8

Table 4

From the above results a number of conclusions can be drawn:

1. The score values from Table 1 are close or may even exceed the previous published results. The reason for higher performance scores stems from the fact that we have computed the highest value from all possible feature combinations while the results from other papers have been obtained as result of running SVM on the full sets of features.

2. As expected, the performance results for 180 training days are lower than the corresponding 30 training days score values (Table 1 vs. 2 and Table 3 vs. 4)

3. No significant performance variations have been noted due to changes in the SVM parameters C and sigma (Table 1 vs. 3 and Table 2 vs. 4)

4. There are no common technical indicator descriptors among securities from the same sector

5. The high Sharpe ratio values (Table 3) are the result of training the SVM engine on a large training dataset while testing on small testing dataset.

These high Sharpe ratio values may be misleading since they are based on results for only 30 or 180 training days. In order to fully characterize the SVM performance a full back-testing process based on a rolling window is required.

We ran additional tests using two SVM rolling window tests on MSFT with C = 100, sigma = 0.071 and technical indicators 1 (EWMA), 3 (RSI), 7(Williams R%).

In a first rolling window test the training sample consisted of 1000 entries with 30 testing days. The iterations are as follows:

Iteration 1:train indices [1:1000], test indices [1001:1030]

Iteration 2:train indices [31:1030]   test indices [1031:1060] …

The P&L graph is shown in Figure 1 below for a Sharpe ratio value of 0.04 (vs. a value of 6.64 in Table 3).
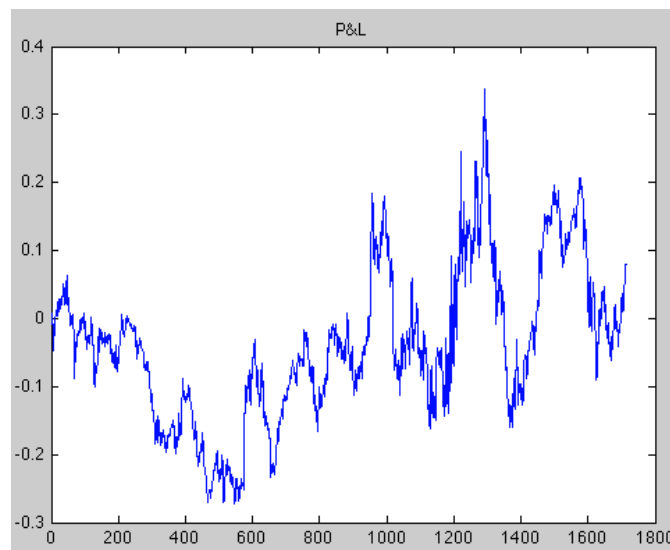

Figure 1

In the second rolling window test the iterations were as follows:

Iteration 1: train indices [1:1000], test indices [1001:1030]

Iteration 2: train indices [1:1030]   test indices [1031:1060] …

The P&L graph is shown in Figure 2 below for a Sharpe ratio value of 0.003.
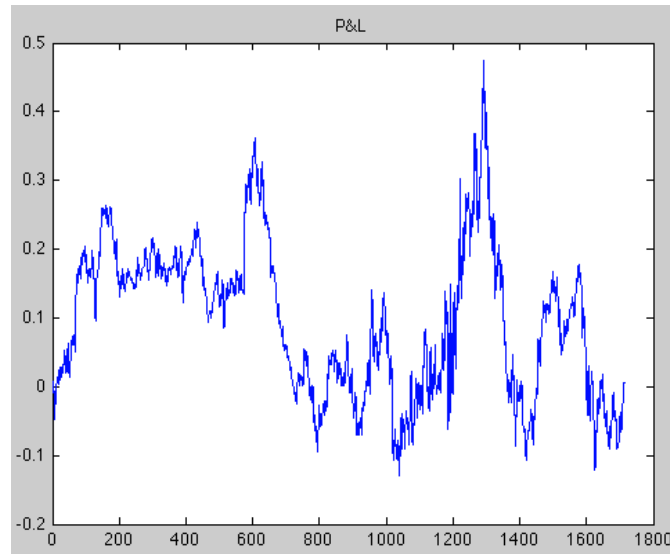

Figure 2

From the rolling window tests we conclude that:

1. SVM models generally fail to predict the market direction of blue chip stocks and S&P 500 index out of sample.
2. Technical indicators based on historical trends, appear to lack predicative power, a finding that is in line with the weak form of the efficient market hypothesis.
3. There is a large difference between the SVM performance results published in the literature and the full back-testing results produced here.

PART 2

In the second research program we examined the ability of different nonlinear classification techniques to predict market direction in the SPY ETF. We used the 11Ants system to construct models across a wide spectrum of nonlinear classification techniques, including SVM, Random Forests, Bayesian prediction, Logistic and Ridge Regression and Nearest Neighbor. The models were estimated using an in-sample data set consisting of 3,011 observations from November 1993 to December 2008, and tested out of sample on a data set comprising 524 daily observations from January 2009 to January 2011. All of the data were standardized by subtracting the in-sample mean and dividing by the standard deviation, also estimated from the in-sample data.

11Ants offers a straightforward and effective means of estimating and testing a very large number of models and model ensembles within the convenient and familiar framework of MS-Excel. In this case

we were able to examine in excess of 1,000,000 such models, running the software on a powerful server for a period of 36 hours.

The 11Ants system evaluates the in-sample performance of each model and model ensemble using a proprietary "quality" index, which is thought to be based on a combination of goodness of fit measures. The models are ranked competitively on the quality scale and the analyst is able to review the relative performance of the top ten models or ensembles, or examine the performance of the best of each different type of model. In this case, the system reported the ten best performing models to be various configurations of Random Forest and SVM, with quality index ranging up to 5.92%. The best performing models in other categories ranked lower in the quality scale, as follows:

- Nearest Neighbor 4.00%
- Logistic Regression 2.26%
- Decision Tree 1.83%
- Logit regression 1.26%
- Bayes 0.00%

11Ants reports the explanatory power of each variable, in the format shown in Table 5 below. It is worth noting that, with the sole exception of the 7-day Exponential Moving Average, the most influential variables are gauged to be the lagged values of the price data, rather than any of the technical indicators.

Variable	Influence
OpenL1	7.7%
Close	7.5%
High	7.1%
EWMA7	6.6%
Low	6.6%
Open	6.1%
EWMA50	6.0%
ADLINE	5.1%
WPCTR	5.1%
EWMA200	5.1%
WADL	4.5%
RSI	4.3%
OBV	3.9%
CHOSC	3.9%
MACD	3.9%
MOM10	3.9%
Volume	3.8%
ACCEL10	3.3%
ROCL1	3.0%
CAMA	2.7%

Table 5

The performance of the top performing model is summarized in Table 6 below. As is frequently the case with non-linear models the in-sample fit is exceptional: the number of correct direction predictions (2,97), representing over 78% of the total sample, is highly unlikely to be the product of random chance. Furthermore, a market-timing strategy using the 1-period ahead forecasts to take a position in the SPY ETF would produce total gross returns of 1839% over the period to December 2008, representing an average annual return in excess of 120%. The equity curve during the in-sample period, shown in Figure 3 below, is smooth and monotonic, with an information ratio of 7.84.

By contrast, the out-of-sample performance is relatively poor: while there is some indication of predictive power, the % Direction Prediction Accuracy is only marginally superior to that of the naïve predictor and is likely to be the product of random chance. This is reflected in the inferior performance of the market timing strategy during the two year period to Jan 2011, which features a drawdown of almost 20%, an information ratio of only 0.52, and flat-to-negative total returns for the thirteen month period from Dec 2009 to Jan 2011.The equity curve during the out-sample period is shown in Figure 4 below. The inconsistency of the results for the in-sample vs. out-of-sample data indicates that the former may simply be the product of curve fitting.

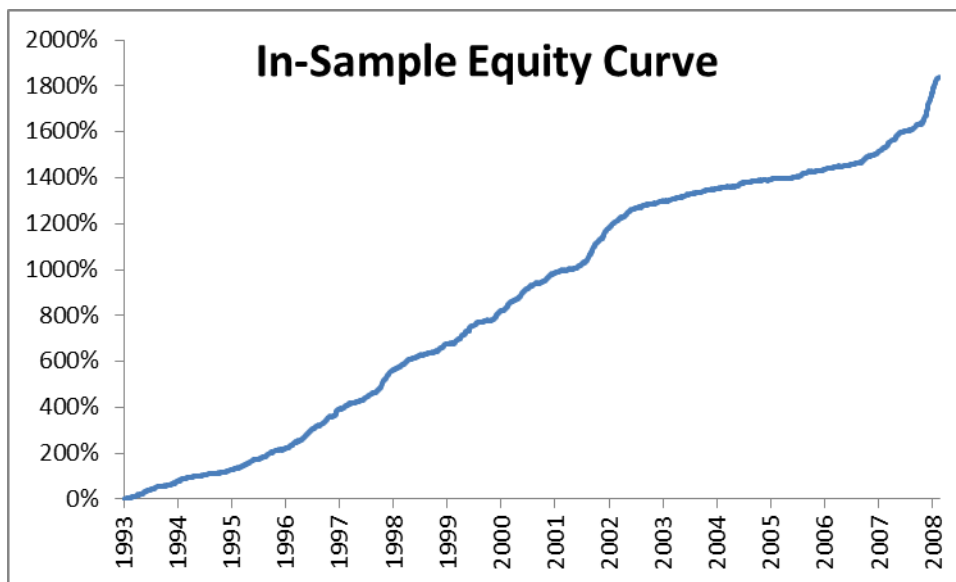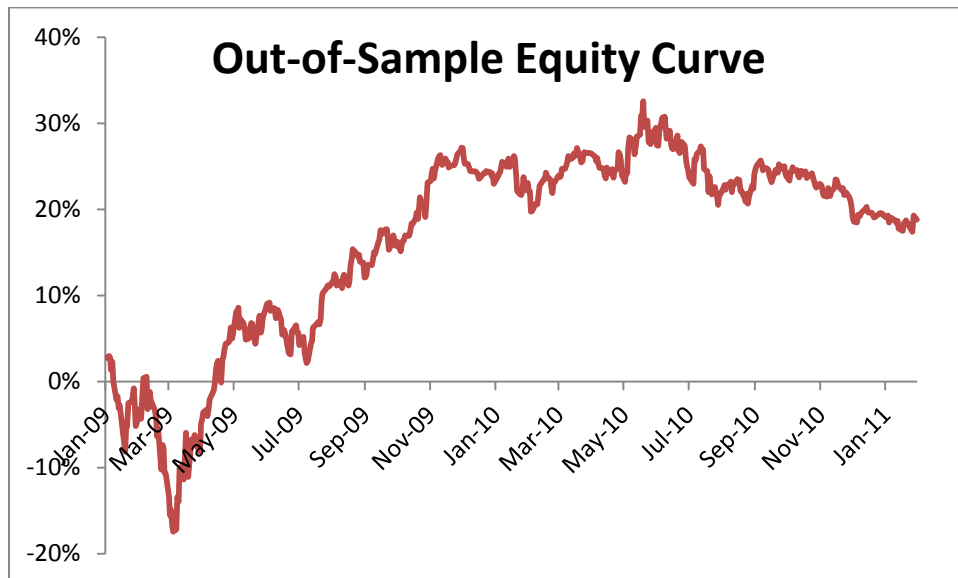|	Nov 93-Dec 2008	Jan 2009 - Jan 2011
# Observations	3810	524
# Correct Direction Predictions	2976	270
% Direction Accuracy	78.10%	51.50%
Total Return	1839.01%	18.81%
Annual Return	120.67%	8.98%
Standard Deviation	15.39%	17.24%
Information Ratio	7.84	0.52

Table 6

Figure 3

Figure 4

4. CONCLUSIONS

In this research paper we have tested the ability of nonlinear machine learning models to forecast market direction using technical indicators as features.

Overall, our research does not corroborate the results reported by some other studies. We find no evidence that SVM techniques enjoy any advantage over traditional methods of forecasting market direction. Furthermore, none of the common technical indicators tested in our study is able to add predictive power to models based on historical price data alone. Consistent with the weak form of the Efficient Market Hypothesis, it appears that all relevant information is incorporated in the current price and neither technical indicators nor non-linear modeling techniques are capable of revealing additional information, at least for the stocks and indices examined here.

An exhaustive test of over 1,000,000 models employing a broad array of non-linear classification techniques produced excellent results in-sample, but failed to perform well out-of-sample. This failure to generalize is a pattern seen in other research and suggests that the in-sample results are the product of curve fitting, something that non-linear classification methods are often able to accomplish very well. The inference is that the positive results reported in some other studies may be the product of limited sample selection.

REFERENCES

Achelis, S.B. , Technical Analysis from A to Z, Second printing, McGraw-Hill, 1995

Breen, William, Lawrence R. Glosten, and Ravi Jagannathan, 1989, Economic significance of predictable variations in stock index returns, *Journal of Finance* 44, 1177-1189.

Cao, L., &Tay, F. (2003), Support vector machine with adaptive parameters in financial time series forecasting, *IEEE Transactions on Neural Networks*, 14, 1506–1518.

Cheung, Yin-Wong, Menzie D. Chinn, and Antonio G. Pascual, 2003, Empirical exchange rate models of the nineties: are any fit to survive?, NBER Working Paper 9393.

Elliott, Graham, and Takatoshi Ito, 1999, Heterogeneous expectations and tests of efficiency in theYen/Dollar forward foreign exchange market, *Journal of Monetary Economics* 43, 435-456.

Gencay, Ramazan, 1998, Optimization of technical trading strategies and the profitability in securitymarkets, *Economics Letters* 59, 249-254.

Huang, W., Nakamori, Y., Wang, S., Forecasting stock market movement direction with support vector machine, *Computers & Operations Research* 32 (2005) 2513–2522

Kim, K*., Financial Time Series Forecasting Using Support Vector Machines,* Elsevier, March 2003

Kuan, Chung-Ming, and Tung Liu, 1995, Forecasting exchange rates using feed-forward and recurrent neural networks, *Journal of Applied Econometrics* 10, 347-64.

Larsen, Glen A. Jr, and Gregory D. Wozniak, 1995, Market timing can work in the real world, *Journal of Portfolio Management* 21, 74-81.

Leung, Mark T., HazemDaouk, and An-Sing Chen, 1999, Forecasting stock indices: a comparison of classification and level estimation models, working paper, Indiana University.

Leitch, Gordon, and J. Ernest Tanner, 1991, Economic forecast evaluation: profits versus the conventional error measures, *American Economic Review* 81, 580-590.

Levich, Richard M., 2001, "*International Financial Markets*", Second Edition, Ch. 8, New York: McGraw-Hill.

Pesaran, M. Hashem, and Allan G. Timmermann, 1995, Predictability of stock returns: robustness and

economic significance, *Journal of Finance* 50, 1201-1228.

Pesaran, M. Hashem, and Allan G. Timmermann, 2001, How costly is it to ignore breaks when forecasting the direction of a time series?, Manuscript, Cambridge University and UCSD.

Rao, S., and Hong, J., Analysis of Hidden Markov Models and Support Vector Machines in Financial Applications, Technical Report No. UCB/EECS-2010-63

Wagner, Jerry, Steve Shellans, and Richard Paul, 1992, Market timing works where it matters most: in the real world, *Journal of Portfolio Management* 18, 86-90.

Wang, L. and Zhu, J., 2008, Financial market forecasting using a two-step kernel learning method for the support vector regression, Ann Oper Res (2010) 174: 103–120

White, Halbert, 2000, A reality check for data snooping, *Econometrica*68, 1097-1126.

Womack, Kent L., 1996, Do brokerage analysts' recommendations have investment value?,*Journal of Finance* 51, 137-167.