Quantitative Investment with Machine Learning in US Equity Market

Yuxiang Huang

International School of Jiangxi University of Finance and Economics, Nanchang, China

Abstract. Quantitative investment attempts to use computer algorithms to predict the price of securities and make automatic trading, in order to gain excess return on the stocks. This paper introduces a strategy based on machine learning algorithms and technical indicators. The model uses several popular technical indicators as inputs and predicts the movement of stock price after a certain short period. Then, a portfolio is constructed using the prediction result. The strategy buys the stocks whose returns exceed the predetermined threshold and sells (shorts) the stocks whose return are below the negative threshold. The empirical results show that the annual return is above 40%,which is far higher than the S&P500 index(2.14%). Considering the risk-adjusted return, the machine learning strategy is better than the S&P500 index. The Sharpe Ratio is higher than that of the S&P500.

Keywords: Quantitative, machine learning, equity market.

1. Introduction

One of the core issues of quantitative investment is to accurately predict the price changes of assets. If you can predict the rise and fall of asset prices, then investment managers can build a more effective portfolio to obtain greater investment returns. Under the traditional analyst-fund manager model, analysts predict the future price of assets through personal subjective analysis, but the model faces two challenges: First, the dramatic increase in the number of assets makes the traditional model to have significant implementation costs, which makes the model no longer applicable; secondly, the analyst's subjective judgment is influenced by his/her emotions and make its predictions no longer accurate. However, the quantitative investment model can solve these two challenges. This paper attempts to construct a quantitative investment model to solve the problem of asset and securities investment price prediction. Machine learning obtains reproducible patterns from large amounts of data and using learned patterns for prediction. Due to the high prediction accuracy and good generalization ability of the machine learning algorithm, the excellent predictive ability of the model can be applied to the off-sample data. This technology is widely used in quantitative investments, such as risk prediction and so on. Technical analysis is widely used by market practitioners to predict future asset prices in actual transactions. Compared with the fundamental data, the technical indicators are easy to obtain from the market data and have strong timeliness. Existing technical analysis studies usually simulate a single technical indicator or technical pattern to verify whether a technology is effective, but the research considering multiple technical indicators or technical analysis methods is relatively lacking. In theory, effectively mining information from multiple technical analysis methods can provide better prediction results than a single technical analysis method. In order to better explore the information in technical analysis, this paper designs a quantitative investment algorithm based on machine learning and technical indicators. The model uses machine learning algorithms to mine a variety of common technical indicators to predict the direction of the stock price after a few days (rise or fall), and then build a portfolio based on the forecast direction. This analysis consists of three strategies, namely neural network, support vector machine,k-nearest neighbors.

1.1 Literature Review

Since Charles Dow put forward Dow Theory in the late 19th century, technical analysis has been widely used in the investment field and actual investment.At that time, evidence showed that technical analysis is feasible for stock market investment. Irwin (1986) established an automated trading system to handle the stock market's complex information. Under the control of the program, the trading system could automatically issue trading signals to provide a reference for the short-term futures funds. Experiments show that nearly half of these signals are valid. Neftci (1991) added the trading signal generated by the 150-day moving average as a dummy variable to the autoregressive equation to analyze the Dow Jones industrial market, which shows that the 150-day moving average trading rule has certain predictability. Ritter(1992), based on behavioral finance, analyzed and confirmed that stocks had a reversal effect, that is, the stocks that have the worst performance in the past few months tend to perform better in the following months. Using the interesting phenomenon, they built a simple quantitative stock picking strategy, sorting all stock returns from low to high on the last day of each month, and then hold the top 10% of the stocks for one month. This is one of the most famous strategies in the field of quantitative investment, namely the reversal strategy. Jegadeesh(1993) found that stock prices have a momentum effect on a larger time scale (3-12 months), that is, stocks that performed better in the past 3-12 months, tend to be robust in the next few months. Based on this phenomenon, they built a momentum strategy that is contrary to the reversal strategy. Lo (2000) proposed a systematic and automatic approach to technical pattern recognition using nonparametric kernel regression and applied this method to many U.S. stocks. With the rapid development of computer science and technology, some methods based on data mining and machine learning, such as neural networks and support vector machines, are also widely applied in the field of stock selection. Kim (2003) used the support vector machine (SVM) to predict the stock price index and discussed the feasibility of using SVM for stock price prediction and made a comparison between this algorithm and BP neural network. The empirical results demonstrated that SVM is a promising stock market prediction algorithm. Khan AU (2008) compared between the accuracy of stock price predictions of some technical indicators, such as back-propagation neural networks, and genetic-based back-propagation neural networks. The results showed that back-propagation based on genetic algorithm has a high accuracy rate for stock price forecasting and should be widely used;Tsai and Hsiao(2010) combined multiple feature selection methods to identify more representative features for the prediction model. The empirical result shows that show that the intersection between PCA (Principal Component Analysis) and GA (Genetic Algorithms) and the multi-intersection of PCA, GA, and CART (Classification and Regression Trees) perform the best and result in 14 to 17 important features respectively. Lee and Jeong (2012) utilize DTW algorithm which had been used for speech recognition and found the utilize of reinforcement learning techniques can successfully handle the risk-averse case and offer a statistical arbitrage strategy for profit even with trading costs. Ticknor et al. (2013) proposed an innovative method: Bayesian regularized neural network to predict financial market behavior, using market prices and financial technical indicators as inputs to predict the stocks of the closing price of the future, empirical results show that the model will perform as well as the advanced model even without data pre-processing, seasonal analysis, and cycle analysis.Patelet al.(2015) found that random forest outperforms other three prediction models on overall performance and that the performance of all the prediction models improve when these technical parameters are represented as trend deterministic data.

1.2 Contribution

The main contributions of this paper are as follows: Firstly, this paper proposes a quantitative investment model based on machine learning and technical analysis. This model creatively uses machine learning technology to classify and predict various technical indicators and builds quantitative investment model based on them. Secondly, the empirical research in this paper finds that machine learning technology can effectively mine the effective information of technical analysis and obtain better investment performance in the stock market. The structure of this paper is as follows: Section 3 sorts and summarizes the current situation of the formation. Section 3.2 presents a quantitative investment model based on machine learning and technical indicators. Section 4 uses the data from the stock market to conduct an empirical study, analyzing the empirical results and conducts a sensitivity test of the model. Section 5 gives the conclusions of this paper and points out the future research directions.

2. Methodology

In this section, I will outline the source of the data and introduce the construction of the technical indicators as features.

2.1 Data

In this research, we used stock daily data which are the constituents of the US S&P500 equity index. The dataset on stocks are downloaded from the Thomson Reuters database. The data contain information on the daily closing(adjusted and unadjusted), opening price as well as their high, low, and trading volume on a daily basis for the period from February 2000 through September 2012.

For each stock and for each year in the sample, this section calculates the measures of return and risk of a certain stock: mean return is calculated to represent the overall performance of each stock throughout the 12-year sample period and the standard deviation of the return, on the other hand, is calculated as a measure of the risks.

We first compute the time-series average of these volatilities for each stock and then report the cross-sectional average of these average volatilities. The other statistics are computed in a similar fashion. For each year and for each stock, the performance is represented using the mean and standard deviation of daily realized stock returns over the most recent 12 months. For each month and for each stock, skewness describes the degree of distortion from a normal distribution in certain set of data. The sample includes 595 stocks and is composed of 3174 monthly returns. The sample period is 2000 to 2012. As an illustration, the descriptive statistics of the Apple Company stock is presented below:

Table 1. The descriptive statistic of the dataset

	Mean return	Volatility	Skew return	Number of observation
2000	1.325	1.025	-6.105	231
2001	0.392	0.622	-0.061	248
2002	-0.424	0.489	-0.588	252
2003	0.400	0.368	0.297	252
2004	1.103	0.396	1.123	252
2005	0.803	0.387	-0.124	252
2006	0.166	0.381	0.703	251
2007	0.851	0.375	-0.086	251
2008	-0.839	0.586	-0.428	253
2009	0.904	0.336	0.210	252
2010	0.426	0.266	0.079	252
2011	0.228	0.262	-0.118	252
2012	0.726	0.269	0.748	176

2.2 Technical Indicators and Features for Machine Learning

In the prediction of the index trend, the collected raw data is so big and abundant that they are not directly used. In this paper, the raw data is transformed into feature data for prediction. A technical indicator offers a different perspective from which to analyze the price action. Some, such as moving averages, are derived from simple formulas and the mechanics are relatively easy to understand. Others, such as Stochastics, have complex formulas and require more study to fully understand and appreciate. Regardless of the complexity of the formula, technical indicators can provide a unique perspective on the strength and direction of the underlying price action. Summarizing previous research, I find that some of the technical indicators are widely used in trend prediction and have good predictive effects.

Table 2. Features construction

	type	function	Representative indicators
1	Leading indicators	designed to lead price movements	Momentum Relative Strength Index (RSI) Stochastic Oscillator Williams %R
2	Momentum oscillators	Measures the rate-of-change of a security's price	Relative Strength Index (RSI) Money Flow Index(MFI)
3	Lagging indicators	follow the price action and are commonly referred to as trend-following indicators	moving averages (exponential, simple, weighted, variable) and MACD.
4	volume	Predict trend using the relationship between volume and price	Accumulation Distribution Line(AD)
5	SSL	Analyze the pressure and support upon the stock price	Bollinger bands(BOLL)

This research selects some representative and important commonly used stocks indicators as input features and variables for the prediction model. They are: Simple Moving Average (SMA), Average Directional Indicator (ADX), Balance of Power (BOP), Money Flow Index (MFI), Standard Deviation (STDDEV), Average True Range (ATR), Momentum (MOM), Bollinger Bands (BB), Chaikin A/D Line (AD), Chaikin A/D Oscillator (ADOSC).

The feature data obtained by characterizing the original data implicitly contains price, time information of the original data sequence. These indicators perform moving average or weighted average processing on the original data, and smooth out the data. As a result, the adverse effects of noise on the prediction model in the raw data can be moderated.

For each stock on each date t, we calculate the technical indicators using the historical data until date t of the stock. At the same time, we calculate the future return of the stock by comparing stock price on dates t+N and t.

2.3 Machine Learning and Prediction

When it comes to stock prediction, we tend to think that the construction of the investment portfolio is mainly based on the future upward and downward movements of stock prices. The research attempts to use the technical indicators calculated above as features for our machine learning algorithms to predict stock return in the next N days. N is considered as the holding period for our portfolio construction which is described in the next Section. Specifically, at any given date t, for each stock individual, we choose the computed normalized technical indicators of the stock from the first available date till t-N as the features for the training dataset. Correspondingly, object variable in the training dataset to train the machine is the future return of the next N days. Of course, when the length of the training set is less than a year at the beginning of the dataset, we do not perform any training nor prediction.

Also, the portfolio construction methods are as follows: Before every single holding period, the investment target is selected based on the prediction model, more specifically, the predicted rate of return. Then, the rate of return will be sorted from high to low, at the same time, a threshold is set. For the stocks that rises more than the threshold, we buy them. For the stocks that falls more than the threshold, we sell or short them.

2.4 Construction Portfolio

Based on technical analysis and pattern recognition, this paper constructs a short-term quantitative investment strategy through machine learning algorithms. After predicting the stock return, this paper buys those stocks whose price is predicted by the machine learning algorithm to rise more than 5% and sells those whose price is predicted to fall over 5% and construct an equally-weighted portfolio.

This stock picking strategy will close these opened positions every seven days (the rolling period), and each day the new position is created through the trading policy and the machine learning algorithms. It is showed by the empirical results that the trading policy mentioned above will outperform the S&P 500 index.

When constructing the portfolio, to simplify the problem, this paper adopts the equal-weighted strategy, that is, for every individual stock in a portfolio, we invest in the same amount of money. After a rolling period, the position that is constructed before such period is closed. So for every single day, we make prediction, select target stocks, build portfolio and close the position which is constructed holding-period ago.

3. Empirical Results

When evaluating the performance of the strategy, this paper adopts most commonly used indicators: Annualized Returns, Volatility, Sharpe Ratio and Max Drawdown. The specific formula is shown in the appendix. The annualized return, which is the annual-equivalent return for the portfolio, measures the profitability of the quantitative strategy, the lager is the annualized return, the better is this strategy. While the max drawdown measures the risk of the strategy. The smaller is the value, the better is the strategy. After calculating the indicators mentioned above, this paper chooses the S&P500 as the benchmark. Not only should we examine the absolute amount of the indicators, but we should also compare the indicators with those of S&P500 index.

3.1 The Examination of the Performance of the Portfolio

In order to evaluate this quantitative trading strategy, this paper will compare the return on the portfolio with the return on the S&P500 index fund. So, it is assumed that at the very beginning of the construction of the portfolio, the same amount of money is invested in the index-driven stock/fund and the price of the stock/fund is purely determined by the S&P500 index. At each end day of the rolling period, the total value of the portfolio is calculated so as to be compared with that of the index-driven fund. When assessing the performance of the portfolio, the important indicator, the growth rate of return, which is mentioned above, should also be taken into consideration. What is more, some risk measures such as Sharpe ratio are also considered when evaluating the performance of the trading strategy. The figure of accumulative profit of the stock ACE is shown below as an example. The profit is equal to the sum of two parts, the cash and the unrealized profit of stock positions.

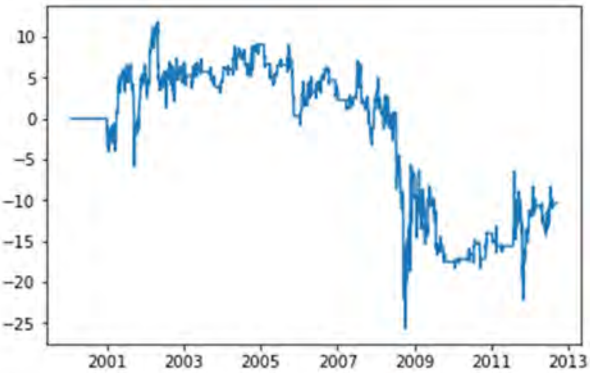

Figure 1. The accumulative profit of a stock ACE is as follows. The profitis measured by the total value of the cash and unrealized profit of stock positions

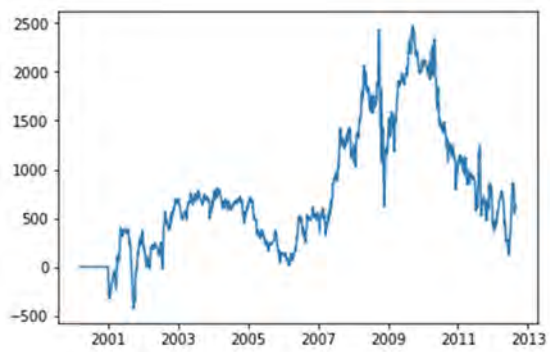

Figure 2. The accumulative profit of the portfolio which is constructed by buying those stocks predicted to rise more than 5% and sells those predicted to fall over 5% and thus forming an equally-weighted portfolio.

The overall performance of the machine-learning strategy can be seen in Figure 2. During most periods, the total value of the portfolio is above zero. The graph shows an obvious upward trend after 2006 and the total value of the portfolio reached its peak of 2500 in 2008 to 2009. After the period, the performance of the portfolio goes down year by year.

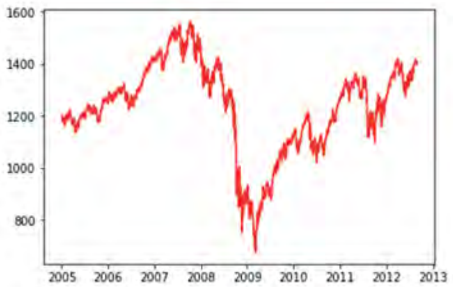

Figure 3. The daily closing price of the S&P500 index.

The above figure shows the performance of the benchmark, the S&P 500 index. The value of the index shows an upward trend between 2005 and 2008. After 2008, possibly because of the sub-prime mortgage crisis, the overall performance is severely weakened and dropped to the bottom of 800 in 2009. Soon after the financial crisis, the depression of the performance recovered and climbed to nearly 1400 in 2013.

The following table shows the absolute return and the risk-adjusted indicators of the strategy and the S&P500 index. As is shown in the result, the return of the strategy based on machine learning strategy is significantly above zero. More specifically, the return on the Neural Network strategy is almost 20 times of that of the S&P500 index, which implies the profitability of the machine learning strategy. Meanwhile, when comparing the risk-adjusted return, we can find that the strategy that is based on the neural network algorithm is more desirable. For example, the Sharpe ratio of machine learning-based strategy is much higher than that of the S&P500 index, meaning that the average return earned in excess of the risk-free rate per unit of volatility or total risk is significantly higher.

Table 3. The comparison of S&P500 index and the strategybased on Neural Network algorithm

Strategy	S&P500	Neural Network Strategy
Annual Return	2.14%	45.07%
Volatility(standard deviation)	0.22	2.72
Sharpe ratio	0.09	0.17
Max drawdown	-888.62	-2346.54

4. Discussions

The results in the last section show that the profitability of the machine learning strategy is better than that of the S&P500 index, which sends us a basic signal that the machine learning strategy is more desirable than the passive investment. In order to check the performance and robustness of the strategy presented above, this section set different scenarios. When different algorithms and holding periods are adopted, the results still show the comparative advantages of the strategy based on the machine learning strategy. The results of each scenarios are shown in the table below, which can lead us to make several discussions as follows:

All the performance listed below are on an out-of-sample basis, whichadopted the rolling window analysis. This paper supposes the training set to be data from the beginning to the date that is seven days before the predicted date. Then, the prediction is made based on each rolling window subsamples.

Overall, the machine learning (SVM and NN) strategy perform better than S&P500 index in terms of the absolute and relative return. In every single group of holding period, the machine learning strategies have an excess return of over 20 times of the index. Also, the risk-adjusted return, which is measured by Sharpe Ratio, demonstrates that the machine learning strategy has relatively more investment potential.

We know that, when the holding period is T=9, the algorithm is the neural network, the return of portfolio is much higher than the other portfolios, and the effect is satisfactory. Specifically, in each group of holding period, the NN strategy outperforms the SVM strategy, which partly shows the relative advantage of the prediction ability of NN than SVM. In each scenario, under the assumption of risk neural, the expected return E(R) represents the required return of the portfolio.

Through the analysis process, the best strategy at present seems to be the T=9, NN strategy. Certainly, more scenarios can be taken into consideration and there are lots of parameters that can be moderated in order to improve the performance of the portfolio. There is still great possibility for us to find better strategy. But limited by the length of this paper, this deeper research can be left as a direction of future studies.

(1)

Table 4. Predictionresultsbasedon different strategies

Holding Period	Strategy	Annual Return	Sharpe Ratio	Volatility
T=5	S&P500	2.14%	0.095	0.2249
	Neural Network	48.32%	0.2682	2.5813
	SVM	44.47%	0.1374	3.2351
T=7	S&P500	2.14%	0.095	0.2249
	Neural Network	45.07%	0.1696	2.7662
	SVM	45.06%	0.1654	2.7237
T=9	S&P500	2.14%	0.095	0.2249
	Neural Network	51.18%	1.0862	0.88024
	SVM	43.86%	0.1586	2.7649

4.1Machine Learning the Prediction Performance and Overfitting.

Neural networks contain multiple non-linear hidden layers, which makes them very expressive models that can learn very complex relationships between their inputs and outputs. With limited training data, however, many of these complicated relationships might be the result of sampling noise, and they will exist only in the training set but not in whole test data even if it is drawn from the same distribution. This leads to the problem of overfitting and many methods have been developed in order to eliminate it. These include stopping the training as soon as performance on a validation set starts to get worse, soft weight sharing (Nowlan and Hinton, 1992) and introducing weight penalties of various kinds such as L1 and L2 regularization.Since the solution of overfitting is not the focus of this paper, it will not be discussed in detail.

5. Conclusion

The purpose of this paper is to predict the direction of the movement of stock price. Prediction model of machine learning algorithms like Neural Network, SVM are applied based on the 12 years (2000 to 2012) of historical data of constituents of the US S&P500 equity index.

When generating features, this paper chooses ten technical indicators to construct the feature base. In order to fully describe the stock market, five groups of technical indicators are all used as input features, they are price indicators, momentum and reversion indicators, stochastic oscillator indicators, volatility and return indicators and volume indicators. These features implicitly contain price, time information of the original data sequence. These indicators perform moving average or weighted average processing on the original data, and smooth out the data. As a result, the noise on the prediction model in the raw data can be moderated. However, other macro-economic variables like exchange rates, inflation rate, government policies etc. that influence stock market can also be adopted as the inputs to the models or in construction of the knowledge base of the quantitative investment system.

The results show that the all the strategy has an excess return compared with that of the passive strategy, which is measured by the performance of S&P500 index. More specifically, the strategy of the NeuralNetworkalgorithm of a rolling period of 9 days outperforms the other scenarios and will have the most desirable return on risk.

In addition, the focus of this paper is short term prediction. In this paper, technical indicators are derived based on the rolling period of 5, 7, 9 days, respectively. It is also worth exploring deeper the significance of the length of holding period and long-term investment can also be a research direction which may involve the analysis of quarterly or annually performance, revenues, profit of companies.

References

[1]. Lukac, L.P., Brorsen, B.W. and Irwin, S.H., 1986. A comparison of twelve technical trading systems with market efficiency implications. Station bulletin-Dept. of Agricultural Economics, Purdue University, Agricultural Experiment Station (USA).

[2]. Chopra, N., Lakonishok, J. and Ritter, J.R., 1992. Measuring abnormal performance: do stocks overreact? Journal of financial Economics, 31(2), pp.235-268.

[3]. Lo, A.W., Mamaysky, H. and Wang, J., 2000. Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. The journal of finance, 55(4), pp.1705-1765.

[4]. Kim, K. J.,2003. Financial time series forecasting using support vector machines. Neurocomputing, 55 (1-2), pp.307-319.

[5]. Khan, A.U., Bandopadhyaya, T.K. and Sharma, S., 2008, July. Comparisons of stock rates prediction accuracy using different technical indicators with backpropagation neural network and genetic algorithm based backpropagation neural network. In Emerging Trends in Engineering and Technology, 2008. ICETET'08. First International Conference on (pp. 575-580). IEEE.

[6]. Tsai, C.F. and Hsiao, Y.C., 2010. Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. Decision Support Systems, 50(1), pp.258-269.

[7]. Takeuchi, L. and Lee, Y.Y.A., 2013. Applying deep learning to enhance momentum trading strategies in stocks. In Technical Report. Stanford University.

[8]. Jegadeesh, N. and Titman, S., 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. The Journal of finance, 48(1), pp.65-91.

[9]. Zivot, E. and Wang, J., 2007. Modeling financial time series with S-Plus (Vol. 191). Springer Science & Business Media.

[10]. Lee, S.J. and Jeong, S.J., 2012. Trading Strategies based on Pattern Recognition in Stock Futures Market using Dynamic Time Warping Algorithm. Journal of Convergence Information Technology, 7(10), pp.185-196.

[11]. Ticknor, J.L., 2013. A Bayesian regularized artificial neural network for stock market forecasting. Expert Systems with Applications, 40(14), pp.5501-5506.

[12]. Ding, X., Zhang, Y., Liu, T. and Duan, J., 2015, July. Deep learning for event-driven stock prediction. In Ijcai (pp. 2327-2333).

[13]. Neftci, S.N., 1991. Naive trading rules in financial markets and wiener-kolmogorov prediction theory: a study of “technical analysis”. Journal of Business, pp.549-571.

[14]. Patel, J., Shah, S., Thakkar, P. and Kotecha, K., 2015. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert Systems with Applications, 42(1), pp.259-268.

[15]. Nowlan, S.J. and Hinton, G.E., 1992. Simplifying neural networks by soft weight-sharing. Neural computation, 4(4), pp.473-493.