Study on Machine Learning Techniques In Financial Markets

Prakhar Vats
B.Tech Computer Engineering
MPSTME, NMIMS University
prakhar.vts@gmail.com

Prof. Krishna Samdani
Computer Engineering Department
MPSTME, NMIMS University
krishna.samdani@nmims.edu

Abstract— Portfolios, securities, stock market forecasting, risk management, debt management are all important pillars of the financial world. These pillars rely heavily on adequate and accurate prediction. These are problems that, on the small scale, affect individuals and their financial conditions and, on the larger scale, can be detrimental to a country's financial well-being. This paper, is an attempt at understanding the various algorithms and platforms involving Machine Learning in Financial Markets and reaching informed conclusions on parameters like accuracy, efficiency, speed and usability. In this paper, primarily, different trading techniques are introduced and their effectiveness in quantitative trading and, in general, finance to generate alphas is observed. These techniques, as observed, are categorised by their reliance on Neural Networks, Support Vector Machines and other quantitative variables in finance. Classifications on the basis of supervised and unsupervised techniques and K-Mean clustering are also made. Further, this paper also delves into the hitherto unpredictable and unmovable phenomena in market and public psychology and attempts to suggest a viable solution to it.

Keywords: Stock Market Forecasting, Machine Learning, Quantitative Trading, Alphas, Support Vector Machines, Supervised Techniques, Unsupervised Techniques, K-Mean Clustering, Market Psychology

I. INTRODUCTION

Prediction has always fascinated the human kind. This fascination when combined with the possibility of financial incentives and the adrenaline of market risk has rendered financial prediction or stock market prediction, in general, of great importance to the world today. Financial markets have evolved drastically with time. In the past, predictions were simply based on available data and qualitative and quantitative assessment by financial analysts. Today, however, with the volatility of markets and the rise of electronic trading platforms, trading analysis and predictions have to range from a few seconds ahead to days or months with data available as well as the data that is getting processed in the present.

This paper focuses on how machine learning methods and models are used to analyse quantitative financial trading to design optimal strategies. The paper also generates insights into how the model limitations brought on by traditional analysis techniques in solving problems involving the voluminous and complex data in today's world are getting eliminated by the use of machine learning to predict prices of stocks, futures and index funds, et cetera.

Naturally, the focus would be on understanding bifurcations of Machine Learning in Reinforcement Learning, Neural Networks, Genetic Algorithms, Decision Trees, Support Vector Machines, Boosting and Expert Weighting. The abilities of these different algorithms in different financial assessments in portfolio management, allocation of funds and financial forecasting in general are also compared on the merits of speed, accuracy, efficiency and ease of use.

Qualitative measures like the psychology of investors and market news are also visited and their influence on market fluctuations are observed. Because of their unpredictable, and sometimes abrupt nature, these predictions are impossible to make using the aforementioned techniques but a viable solution involving image recognition and machine learning is also visited and suggested.

II. LITERATURE REVIEW

A. *Neural Networks*

Neural Networks, in recent times, have become powerful tools to predict and forecast financial markets owing to their capability of nonlinear approximation and their power to process mixed and disordered data. [1] Neural Networks do not need to build explicit relations of complex nonlinear systems and mathematical models. The first research on prediction of stock prices using NN's was in White's Trial in 1988 to predict IBM's daily stock return. The forecast, however, wasn't ideal owing to incorrect selection of activation functions but the research showed promise. Since then, similar trials flourished across the globe; the most notable of which was Tokyo Stock Exchange Price Index to predict prices of weighted average indices in the Tokyo Stock Exchange.

Early results showed that [1] Neural Networks were 19% more accurate in forecasting the price of weighted average

@IEEE 978-1-7281-1524-5

indices for the TSE over traditional forecasting techniques. The neural network used here had 15 input variables, 2 hidden layers, and 1 output variable to predict the stock price trends in Japan.

The training samples were divided into two trends- rising and downward. Although the samples did reasonably well in predicting the upcoming market trends, they weren't strong enough to forecast the future prices of indices. Therefore, NN in the late 80's were primarily used to determine whether the investors should call "long or short" on the stock prices. In 1999, Pesaran, Hashem and Timmerman predicted the London security index for the past 25 years and it had a 60% success rate in anticipating the change of the index per month.

In summary, following drawbacks of NN's were found in prediction of non-linear stock market data.

1. Predictions made by Neural Networks were computationally expensive
2. When dealing with small samples, NN's were susceptible to overanalysing data and sinking into local minimums
3. Generalisation abilities of NN's were not ideal

B. Support Vector Machines

Support Vector Machines are considered to be one of the most suitable algorithms available for time series prediction. This supervised algorithm can be used in both, regression and classification. SVM's benefit from two ideas: maximizing the margin and the kernel tricks. SVM's, using the maximum margin, construct and optimal separating hyper-plane. The original input feature space is mapped into a higher dimension so that it can be separated by a linear model in the higher dimensional space (hard margin). Linear, RBF and sigmoidal kernels can be used for this purpose. Input vectors which define the width of the maximum margin are called support vectors.

In the case of a feature space that is not separable, misclassification of some points is allowed by using a soft margin. A penalty parameter C (C>0) can be set on the upper bound to control the amount of tolerable deviation.

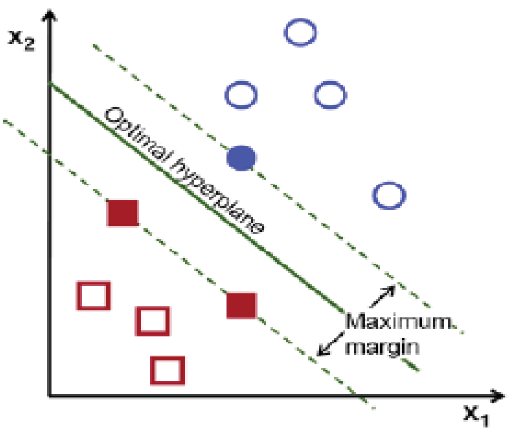

Fig. 1. Geometric desciption of SVM margins

Because of financial markets' complex characteristics of dynamic nonlinearity, non-stationary and low signal to noise ratio, Self-Organising Feature Maps are used to perform cluster analyses on futures, bonds and stock index time series data first. Then selected key features are subjected to SVR (Support Vector Regression). This results in a decreased training time and increased forecasting accuracy.

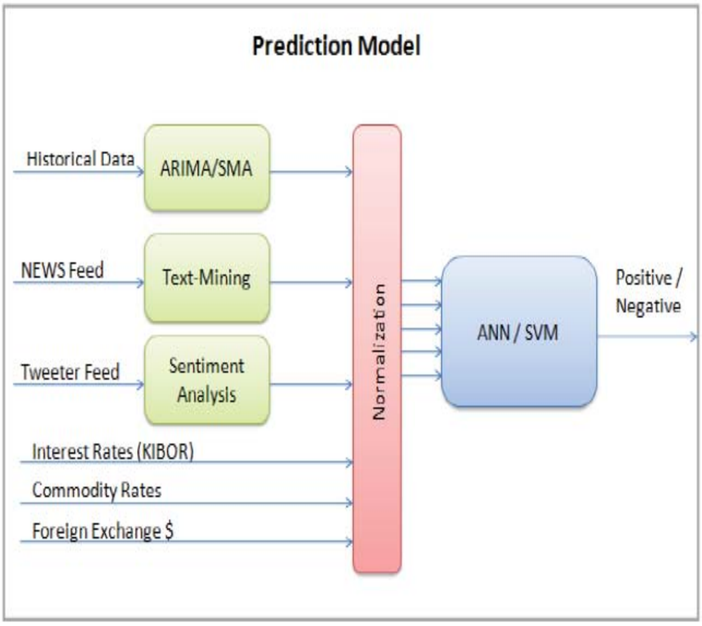

Fig. 2. Prediction Model for effective classification using SVM's [5]

C. Multiple Kernel Learning

[3] Multiple Kernel Learning involves linearly combining fixed base kernels to construct an effective kernel model. The success of SVM depends on the choice of a good kernel, which are prepared in the aforementioned fashion. Learning datasets are used by MKL kernels. In the multiple kernel framework, weighted linear combination of M base kernels is used in manufacturing the optimal kernel.

Equation below shows the process of linear combination of K base kernels.

$$K_{comb}(x,y) = \sum_{j=1}^{K} \beta_j K_j(x,y)$$

$$with\ \beta_j \geq 0, \sum_{\beta_j}^{K} \beta_j = 1$$

(1)

Where Beta of J is the variable for weights assigned to combine base kernels. It is also used to combine different sub-kernels. Using training data, optimal weights to be assigned can be calculated with the use of Multiple Kernel Learning.

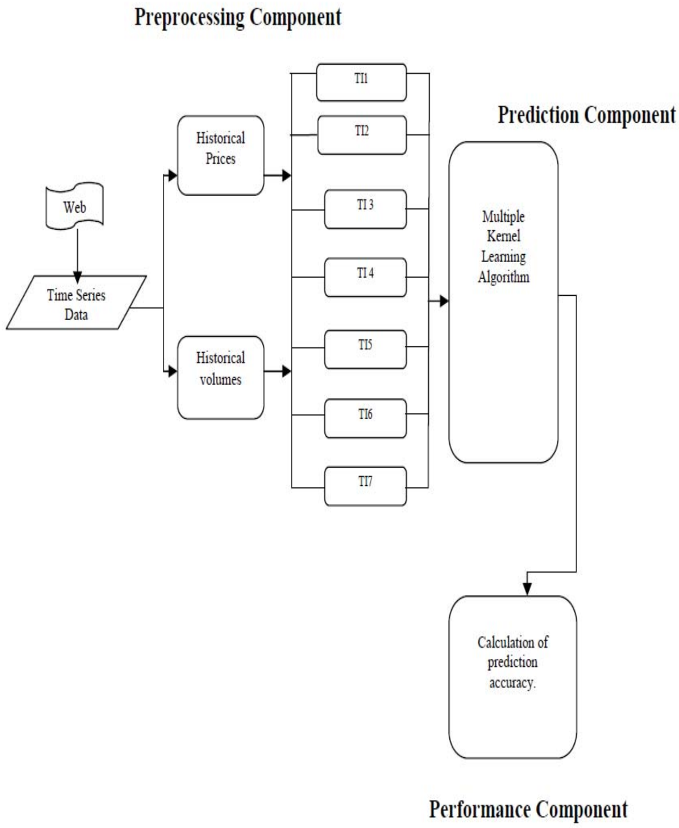

Fig. 3. Proposed Prediction Model for Multiple Kernel Learning [2]

D. Random Forest Learning Method

Random Forest is an ensemble learning method for classifications and regressions problems. [4] It is a combination of decision trees growing in randomly selected feature subspaces and bagging methods to produce data sets for the trees. The results through Random Forests are obtained by classification and regression.

The building strategy for Random Forests is as follows:

1. At every node, input variables are randomly chosen and the best split is calculated within the subset itself
2. All trees in the forest are maximal trees
3. No pruning involved

Random Forest Technique is accurate and stable. The variance of the obtained result is less and there is little to no room for over-fitting

E. K-Means Clustering

K-Means Clustering is an unsupervised technique used mainly for financial forecasting. [5] This MATLAB based clustering technique feature weight elimination by normalisation of attributes to obtain an ideal result.

The normalisation technique is as follows

$$\frac{X - X_{min}}{X_{max} - X_{min}}$$

(2)

K-Means Clustering also involves Principle Component Analysis which features selection of only those components which have a profound effect on the classification. The remaining components are eliminated.

The basic steps of K-Means Clustering are as follows:

1. Initialise centroids that are to be the same as number of classes (targets) in data
2. Determine the distance of each object to the centroid
3. Select minimum distance of the obtained distances
4. Group objects on the basis of minimum distances

Continue steps 3 and 4 until one of the targets is reassigned

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

Fig. 4. K-Means Clustering Accuracy Averages [6]

@IEEE 978-1-7281-1524-5

III. RESULTS AND INFERENCES

	Accuracy	Speed	Efficiency	Ease of Use
Neural Networks	20% initially but modern day accuracy close to 60%	Powerful and speedy due to strong nonlinear approximation	Time complexity given by BIG Oh of Ntrees*N*(log(n)/p)	Easy to moderate for a professional to use.
Support Vector Machines	Best case testing accuracy in the case of Linear SVM found out to be close to 89%	Linear SVM found to be one of the fastest techniques unlike Radial and Sigmoid SVM's	[6] CPU run time less than one cycle for Linear SVM but more than 250 for Radial and Sigmoid SVM's	Supervised technique, so is complex to implement but easy to use for the user.
Multiple Kernel Learning	Accuracy for N Data and S Data both closely matched at 55% to 58%	Speed proportional to the number of predefined kernels	Dependent on system configuration. A 2.4GHz processor and 4GB RAM sufficient for efficient forecast	Sound technical software and financial knowledge required.
Random Forest Method	Depends on test case variables. Best case accuracy observed to be 95%	Dependent on number and size of neurons but considered generally fast	Time complexity depicted by Big Oh of N(neurons)*Size(neurons)*(n/p)	Easy to implement because of predefined quantitative variables
K-Means Clustering Technique	Average case accuracy found to be around 75% and best case found to 79.1%	Proportional to number of nodes but still very fast due to simple quantitative base	Directly proportional to the number of nodes and the number of comparisons	Easy to implement due to a simple quantitative base

IV. PROPOSED FUTURE RESEARCH

One key area that affects market fluctuations is [7] market or investor psychology. This is called "Behavioral Finance." Although we have gotten more and more efficient in predicting market trends using quantitative methods and Artificial Intelligence, market psychology still continues to confound prediction models which have hitherto never factored it in on the same level at which it affects the market.

A suitable solution to using market psychology to our benefits is through monitoring of public interaction with and on e-platforms of financial trading. This would include text mining and Machine Learning to establish a scale of investor emotions which might directly or indirectly affect the market. The proposed model would include constant monitoring of comments and interactions by investors and would be incorporated into a rating scale which would depict a factor by which the investors are "long or short" on the current market.

V. CONCLUSION

With the data presented in the paper, and all the observations and inferences made, it is suffice to assert that computation, analysis and prediction have come a long way in the modern age due to the use of Machine Learning. We have, to ourselves, a plethora of algorithms which can provide to us a varying range of accuracies and efficiencies best suited for our datasets. Keeping the complexity of financial markets in mind, in that, the juxtaposition in preferences of cash and derivative markets in real time trading by investors. Both cash and derivatives markets are closely related for these algorithms to work with adequate efficiencies but in order for us to obtain excellent results in both these markets, we need to highlight the minor differences in these both and have separate methodologies for obtaining optimal predictions and forecasts. In recent times, the aforementioned techniques in Neural Networks and SVM's have seen detailed and unique bifurcations in Genetic Algorithms and Back Propagations along with the Time-Series Wavelet Analyses which further the abilities of these techniques to generate more precise and advanced insights into Financial Markets of the modern day.

With the inclusion of Machine Leaning in Financial Markets, we've observed both- the investing part and the curating part of stock markets and financial markets, in general, get better.

VI. REFERENCES

[1] Yelin Li, Junjie Wu, and Hui Bu, "When Quantitative Trading Meets Machine Learning," 13th International Conference on Service Systems and Service Management (ICSSSM), June 2016.

[2] Amit Kumar Sirohi, Pradeep Kumar Mahato, and Vahida Attar, "Multiple Kernel Learning For Stock Price Direction Prediction," International Conference on Advances In Engineering & Technology Research (ICAETR), August, 2014.

[3] Badre Labiad, Abdelaziz Berrado, and Loubna Benabbou, “Machine Leaning Techniques For Short Term Stock Movements Classification For Moroccan Stock Exchange,” in 11tf International Conference On Intelligent Systems: Theories and Applications (SITA), October,2016.

[4] Mehak Usmani, Syed Hasan Adil, Kamran Raza, and Syed Saad Azhar Ali, in “3rd International Conference on Computer and Information Sciences (ICCOINS),” August,2016.

[5] Zhen Hu, Jie Zhu, and Ken Tse in “6th International Conference On Information Management, Innovation Management And Industrial Engineering,” November, 2013.

[6] Nicole Powell, Simon Y. Foo, and Mark Weatherspoon, in “40th Southeastern Symposium On System Theory (SSST),” March 2008.

[7] Ye Shi-Qi and Peng Yong, “The Relation Between Risk And Return Of Portfolio Based On Standard Finance And Behavioral Finance,” IEEE International Conference On Control And Automation, June 2007.