A Comparison Study of Support Vector Machines and Multi-layer Perceptron Using Daily News for Stock Market Prediction

Junru Xiong

Abstract

New for stock movement prediction is an emerging area in artificial intelligence (AI). This paper will use Support Vector Machine (SVM) and Multi-layer Perceptron (MLP) algorithms to implement stock market movement prediction. It will also use Grid search and Bayesian hyperparameters optimization techniques to optimize those models to generate the best predictor and critically evaluate their performance.

Introduction & Motivation of the problem

Daily news for stock market prediction is currently an active research topic, a longstanding challenge that spurs interest in natural language processing (NLP) and deep learning among academics and practitioners (Krysovatyy et al., 2019). It is a well-established fact that stock prices are driven by market sentiment. It is also reasonable to hypothesize that news has a role in shaping this sentiment. Apart from the obvious benefits to the investors and the analyst community, this technique has the potential to be of immense utility to firms to protect the interests of their shareholders.

While this field has attracted significant interest, a model with reasonable accuracy still evades the seekers. This report aims to use Support Vector Machine (SVM) and Multi-layer Perceptron (MLP) models to predict the stock market's movement using daily news text data from Twitter and optimize their hyperparameters to evaluate their performance and make a brief conclusion critically.

The dataset description, initial analysis and data engineering

This stock news dataset comes from Kaggle (2020), gathered from multiple Twitter handles regarding economic news. The dataset has 2 columns and 5791 rows. The first column represents the news header text, the second column is the stock market close state, where '1' is rose or stays the same compare with the day before, and '-1' is decreased. From figure 1 the number of the stock rose to take up 63.6% (3,685) and fall occupy 36.4% (2,106) in the dataset. Since the difference in the number of samples is not significant, techniques such as SMOTE to equilibrate samples are not necessarily compulsory (Chawla et al., 2002).

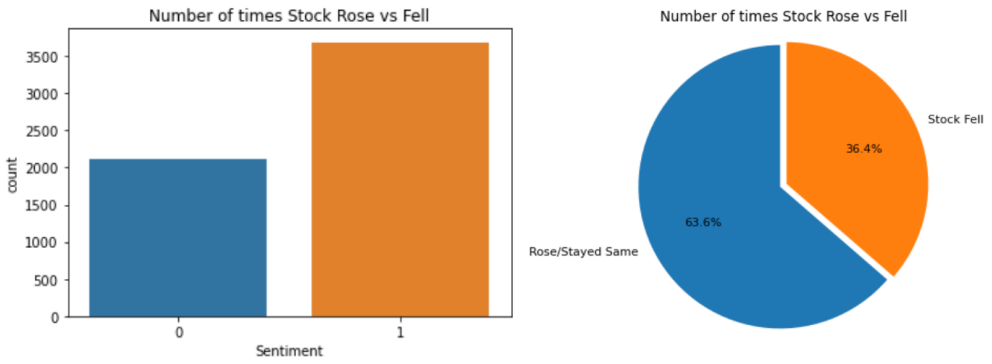


Fig.1 raw data distribution and proportion

To be sure that the text data can be understand by machine, it removed the unnecessary symbols and stop words, align every characters as lower case, lemmatize different forms of words. In this report, it also use 'CountVectorizer' from Sklearn package to convert the one dimension feature text to a matrix of token counts and set a reasonable max feature as 5000. However, as the features' dimensions increase, high-dimensional data sparseness problems may arise (Birnbaum et al., 2013). It may take the training results with limited data over-fitting and difficult to converge as well as make the computing expensive. Therefore, it will use Principal component analysis (PCA) to reduce the data's dimensionality and finally choose 256 dimensions with minimal loss.

Summary of the two models

- *Support Vector Machine (SVM)*

SVM is used for both regression and classification models. Its basic form is a binary linear classifier with the most significant margin defined in the feature space. The learning strategy of SVM is to maximize the soft margin. SVM can be used for a non-linear separable problem by different kernels such as polynomial, Gaussian radial basis functions, sigmoid, which takes the samples to an additional dimension. The dataset can then be divided by a hyperplane (Cortes and Vapnik, 1995). The below table shows the advantages and disadvantages of SVM:

Pros of SVM	Cons of SVM
• Memory efficiency whilst training is very good as SVMs use a few data points to compute the decision boundary.	• The performance of SVM can be drastically reduced if datasets contain many noisy data points.
• It efficiently work in both high dimension sample and small datasets..	• SVM does not directly provide probability estimates, which is calculated by expensive cross-validation
• Training SVMs can be slow for certain kernel functions and large datasets	• If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial.

Fig.2 SVM's pros and cons table

- *The Multilayer Perceptron (MLP)*

The Multilayer Perceptron (MLP), also known as Feedforward Artificial Neural Network, is the most adopted technique for both classification purposes. MLP is a supervised learning method that consists of an input layer, output layer, and few hidden layers. Every neuron is associated with weights, activation function, and a bias. When training the data, it uses the backpropagation technique where the predicted output is compared with the target output, and a loss is calculated at the output layer (Haykin, 2009). This loss will be sent back through the MLP to update the weights of layers, such as gradient descent. This process will reduce loss until it reaches a certain number of training epochs. The below table shows the advantages and disadvantages of MLP::

Pros of MLP	Cons of MLP
• It can learn non-linear models in real time	• It is sensitive to feature scaling.
• Different forms of regularisation to deal with overfitting and improve the convergence such as L1/L2 regularization, early stopping, drop-out and so on.	• It is a black box model which makes it hard to explain the reason of the network produced certain output.

• It is high performing models in supervised tasks especially where the dataset is large, high dimensional, and unstructured.	• The MLP with hidden layers contains a non-convex loss function with more than one minimum, so different random initial weights may lead to different validation accuracies
• it does not make assumption regarding the underlying probability distributions	• It takes flattened vectors as inputs, which disregards spatial information

Fig.3 MLP's pros and cons table

Hypothesis

SVM and MLP models performances according to the tasks they are applied to, and MLP outperforms SVM in many cases. (Frias-Martinez and Velez, 2006). However, Zanaty (2012) states that SVM is typically more suitable on high-dimensional datasets than MLPs. Since this research dataset is 256 dimensions, it assumes that the SVM will outperform the MLP. SVM may also have sensitive performance using a different kernel.

Methodology

Because the raw dataset does not have a time feature, it consists of randomly split 20% of the original dataset as testing data for final model performance evaluation before the modeling process. The rest of the 80% data is used for training, validating, and hyperparameter tuning in model selection and used for training in comparing both algorithms.

- *Architecture and Parameters used for the SVM*

It investigated five different kernel functions of SVM, linear, Gaussian, Sigmoid, and Polynomial to map the non-linear separable data-set into a higher dimensional space to find a separatable hyperplane, and SVM prediction has to vary the sensitivity to the different kernel. It will choose significant parameters such as regularization parameter 'c' (controls the trade-off between smooth decision boundary and classifying training points correctly), five different kernels, degree of the polynomial kernel function, gamma (coefficient for 'rbf', 'poly' and 'sigmoid' kernel) for hyperparameter tuning using both grid search and Bayesian hyperparameter optimization with ten-fold loss error function. It will use the confusion matrix, F-1 score, recall, precision score, and AOC AUC curve to evaluate its performance on the training set when it computes the optimal parameters.

- *Architecture and Parameters used for the MLP*

MLP usually initializes at random weights to avoid to drop in local minima. However, because of its random initialization, the different training may cause results to vary. This study will implement one hidden layer MLP, set the maximum number of epochs to 100, and apply an early stopping criterion to avoid overfitting. It will choose the learning rate, number of hidden layer neurons, dropout, batch size, and momentum for hyperparameter optimization using a 10-fold cross-validation grid search. It also set the SoftMax activation function to solve the classification problem, which returns a probabilistic distribution of the classes.Using confusion matrix, F-1 score, recall, precision score, and AOC AUC curve to evaluate its performance in the training set.

Results, Findings & Evaluation

- *model selection*

Below table 4 shows the grid search hyperparameters and the top 20 models selected for MLP and SVM. In the SVM validation phase, the highest ten-fold validation score is 74.3%, as shown by the first row. The 'rbf' kernel performs significantly well, the best gamma value is

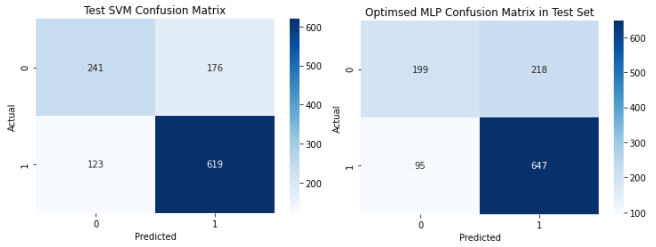
0.1, and the best regularization rate ‘c’ is 10. It confirms the previous assumption that SVM can be sensitively affected by the different kernel, which the lowest validation score is 65.5% when it is using the sigmoid kernel. The best MLP hyperparameters are 128 batch size, 0.05 learning rate, 0.9 Dropout, 128 neurons in the hidden layer, and 0.9 momentum rate. Its best model 10 fold validation accuracy is 73.9%, which has a similar performance as SVM.

SVM					MLP					
C	Degree	Gamma	Kernel	CV Accuracy	Batch size	lr	Dropout	Neurons	Momentum	CV Accuracy
10	NA	0.1	rbf	0.7433	128	0.05	0.9	128	0.9	0.7394
1	NA	0.1	rbf	0.7363	32	0.1	0.5	128	0.9	0.7394
10	2	1	linear	0.7299	64	0.1	0.5	64	0.5	0.7386
10	2	0.1	linear	0.7299	128	0.1	0.5	32	0.95	0.7386
10	2	0.01	linear	0.7299	32	0.1	0.1	128	0.3	0.7381
10	2	0.001	linear	0.7299	32	0.01	0.5	128	0.95	0.7379
10	3	1	linear	0.7299	128	0.1	0.5	128	0.5	0.7375
10	3	0.1	linear	0.7299	128	0.05	0.5	32	0.9	0.7373
10	3	0.01	linear	0.7299	128	0.05	0.5	128	0.9	0.7373
10	3	0.001	linear	0.7299	128	0.05	0.5	128	0.95	0.7373
10	4	1	linear	0.7299	64	0.05	0.9	128	0.9	0.7370
10	4	0.1	linear	0.7299	128	0.1	0.1	128	0.3	0.7370
10	4	0.01	linear	0.7299	128	0.01	0.5	32	0.95	0.7366
10	4	0.001	linear	0.7299	128	0.01	0.5	64	0.95	0.7366
1	2	1	linear	0.7298	64	0.1	0.1	64	0.5	0.7366
1	2	0.1	linear	0.7298	32	0.01	0.5	64	0.95	0.7364
1	2	0.01	linear	0.7298	32	0.05	0.5	32	0.95	0.7364
1	2	0.001	linear	0.7298	64	0.1	0.1	128	0.5	0.7364
1	3	1	linear	0.7298	32	0.05	0.5	64	0.95	0.7364
1	3	0.1	linear	0.7298	64	0.05	0.5	64	0.95	0.7362

Table 4 Top 20 Hyperparametes Grid Search Result

- *Algorithm Comparison*

The model selection process applies a test dataset on the best hyperparameter of SVM and MLP models. The metrics table shows (fig.5) that the test accuracy of SVM is 74%, and the test accuracy of MLP is 73%. F1 score provides a weighted average of precision and recall, and MLP’s F-1 score lower than SVM. So SVM overall performs better than MLP. It validated our assumption in the hypothesis part that SVM outperforms in high dimension classification. It can also be observed from the confusion matrix that the MLP classify much less well in False Negative than SVM so that SVM may lead investors to less financial loss in the stock market.



Metrics	SVM	MLP
Accuracy	0.74	0.73
Precision	0.74	0.72
Recall	0.74	0.73
F1-score	0.84	0.81
ROC-AUC	0.71	0.67

Fig.5 Test set confusion matrix (left), key metrics in test set (right)

A ROC (Receiver Operator Curve) was plotted to enhance this experiment analysis in figure 6 (left), SVM performed slightly higher 4% than MLP, and they are both good performance for stock prediction, which are much higher than 50% AUC (Kalyani, Bharathiand Jyothi, 2016). Figure 6 (right) shows the learning curves of both models. They only slightly converged at 3,500 training data points with high variance, and MLP has a more significant error interval. Potentially, a less advanced text processing method that makes the machine cannot fully

understand the meaning of words or a small quantity of dataset may cause a learning curve unconverging.

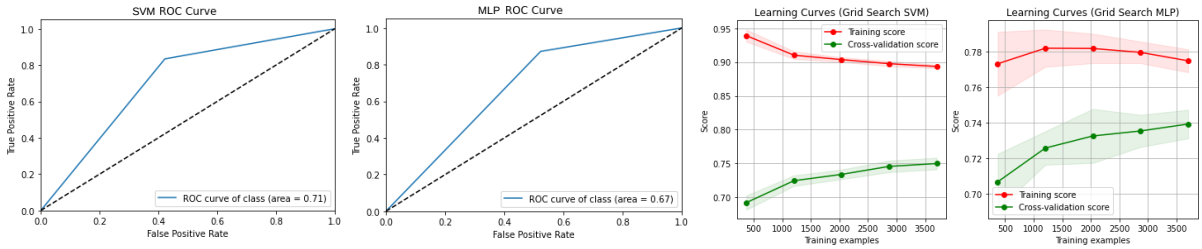


Fig.6 SVM and MLP ROC curve (left), Learning Curves in test set (right)

conclusions, lessons learned, references and future work

Through this paper, it reported the performance of two types of models, SVM and MLP. Throughout this report analysis, it identified the significant sophistication of both models' architecture, their pros and cons, and evaluative performance. Both SVM and MLP models perform well on stock prediction situation and SVM slightly outperform than MLP. The SVM outperformed the MLP during training and testing, which was expected as stated in the hypothesis part; metrics, such as recall, calculated from confusion matrices and ROC plots, proved robust evaluation measures. However, since this data has no time feature, the training set and the test set cannot be split in chronological order, so these prediction results may not fully represent future stocks' trends. It either needs a much larger dataset, time features, or advanced natural language processing method to fully understand the news's meaning, such as the Word2vec technique. Furthermore, investigating unsupervised deep learning, such as Self Organizing Maps (SOM), could potentially prove to be more powerful than our current supervised learning approach for classifying whether the stock fell or rose.

References:

Birnbaum, A., Johnstone, I.M., Nadler, B. and Paul, D., 2013. Minimax bounds for sparse PCA with noisy high-dimensional data. *Annals of statistics*, *41*(3), p.1055.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, pp.321-357.

Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine learning*, *20*(3), pp.273-297.

Frias-Martinez, E., Sanchez, A. and Velez, J., 2006. Support vector machines versus multi-layer perceptrons for efficient off-line signature recognition. *Engineering Applications of Artificial Intelligence*, *19*(6), pp.693-704.

Haykin, S.S., 2009. Neural networks and learning machines/Simon Haykin.

Kalyani, J., Bharathi, P. and Jyothi, P., 2016. Stock trend prediction using news sentiment analysis. *arXiv preprint arXiv:1607.01958*.

Krysovatyy, A., Vasylchyshyn, O., Desyatnyuk, O. and Galeshchuk, S., 2019. News feed for stock movement prediction. In *CEUR Workshop Proceedings* (pp. 90-97).

Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine learning*, *20*(3), pp.273-297.

Zanaty, E.A., 2012. Support vector machines (SVMs) versus multilayer perception (MLP) in data classification. *Egyptian Informatics Journal*, *13*(3), pp.177-183.