Stock Forecasting ML Project based on SVM

Chao Zhou, Yuan Wang

1. Abstract

In the financial field we can distinguish two different approaches: Technical and Fundamental Analysis. The goal is to identify patterns in order to predict time-series movements and improve accuracy. The technical approach is based on analyzing stock market considering previous observed patterns with the aim to determine future time series values. In fundamental analysis (FA), the basic idea focuses on evaluating intrinsic value (i.e. using Discounted Cash Flow Model to calculate a company's present value). Technical analysis (TA) gained considerable attention due to the efficient market hypothesis which claims that stock movements are not a stochastic process but reveal repeated patterns over time. Generally, Technical analysis is the study of historical data and then draws conclusions from the historical market patterns based on the past movement trend.

2. Introduction

In this project, for the purpose of forecasting stock movement effectively, we propose to use the classification method SVM (Support Vector Machine) to improve investors' investment ability and reduce investment risk. Compared with other nonlinear prediction methods, this machine learning algorithm has the advantages of good generalization ability, strong robustness and high prediction accuracy. Our experiment process involves standardizing the stock data, finding the main control factors through GRA (Gray Relational Analysis), dividing the samples into test set and training set, and applying grid search to find the optimal parameters to improve the prediction efficiency. Finally, the empirical results and research conclusions are given, which have some practical guiding significance for the effective prediction of stock data.

3. Background

Stock forecast involves the economic interests of investors, and how to conduct more accurate stock forecasts has been a hot issue of concern. According to the research of experts and scholars at home and abroad, the contemporary financial data show more and more non-linear characteristics, and the traditional time series prediction methods, such as AR, MA, ARMA and other models based on linear data,

have already failed to meet the needs. Although the neural network method can provide more accurate results, it is difficult to choose the network model and structure, with higher possibility of falling into local minimum, slow training speed, poor generalization ability and other phenomena. In recent years, SVM is an effective method to solve nonlinear problems after neural networks, which belongs to the category of machine learning and can overcome many shortcomings of traditional statistical methods and neural networks.

4. Experiments & Methods

4.1 Data Description

We accessed our stocks dataset from https://hk.finance.yahoo.com/, and it is open for downloading and experimenting purpose. The time period selection for our data is from November 2013 to November 2020, and the stocks are respectively the Shanghai stock index and Baidu stock index data, and the following figure is an example given for the graphical representation of Shanghai stock index:

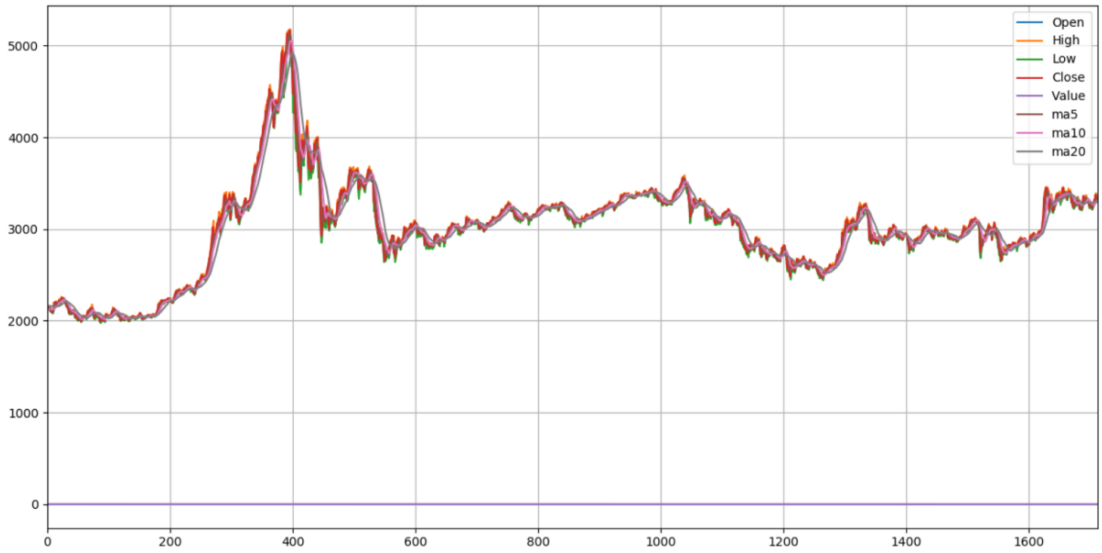

Figure 1. Shanghai Stock index Graphical Representation

4.2 Exploratory Data Analysis

For classification, label data is needed as the dependent variable of the algorithm, so we need to choose by ourselves. Intuitively, we decide our label data by using the today's closing price to minus yesterday's closing price: if the value is greater than 0, it is set as 1 to represent a rising trend; if it is less than 0, it is set as 0 to represent a falling trend. During the analysis of the date, we also found that the value of volume data is very large, which suggests us that standardization is needed. Moreover, by observing the data, we can see that Close and Adj Close is basically the same, so we abandon the feature Adj Close for simplification. Lastly, through utilizing python's third party package Talib, we created EMA (Exponential Moving Average) data of 5, 10, 20 and set them as features as well.

4.3 Pre-processing & Feature Extraction

Data Normalization: Data normalization is scaling the data so that it falls into a specific interval. Different data often have different dimensions, which will affect the results of data analysis. In order to eliminate the dimensional influence between data, standardization should be carried out to solve the comparability issue between data indicators. After standardized processing of the original data, each index is in the same order of magnitude, which is suitable for comprehensive comparative evaluation. The two comparative advantages brought by data standardization:

1) It speeds up the solution optimization of gradient descent;
2) It has the potential to improve accuracy.

PCA (Principal Components Analysis) : It is to project the original sample data into a new space, which is equivalent to the mapping of a set of matrices to another coordinate system in matrix analysis. A transformation coordinate can also be understood as a transformation from one set of coordinates to another set of coordinates, but in the new coordinate system, the original one does not need so many variables, only the coordinates of the space corresponding to the eigenvalues of the largest linearly independent group of the original sample are needed.

4.4 Feature Selection

Gray Relation Analysis (GRA) is a multi-factor statistical analysis method. Simple speaking, in a gray system, we want to know the relationship between index and factors, and rank their order based on their relative correlation strength. Here are the specific steps we took under GRA:

1) Establish the parent sequence;
2) Dimensionless;
3) Calculate the gray correlation coefficient based on the following formula:

$$\zeta_i(k) = \frac{\min\limits_i \min\limits_k \left|x_0(k) - x_i(k)\right| + \rho \cdot \max\limits_i \max\limits_k \left|x_0(k) - x_i(k)\right|}{\left|x_0(k) - x_i(k)\right| + \rho \cdot \max\limits_i \max\limits_k \left|x_0(k) - x_i(k)\right|}$$

Figure 2. Formula for correlation coefficient

4) Calculate the mean value of correlation coefficients and form the correlation sequence:

$$r_i = \frac{1}{n}\sum_{k=1}^{n} \xi_i(k)$$

Figure 3. Formula for value of correlation

In essence, GRA algorithm provides a method to measure the distance between two vectors. For time-dependent factors, the vector can be regarded as a time curve,

while GRA algorithm measures whether the shapes and trends of two curves are similar. In order to avoid other interference and the influence of protruding morphological characteristics, GRA first did normalization, corrected all vectors to the same scale and position, and then calculated the distance of each point. Finally, through the correction of Min and Max, the final output result falls between value 0 and 1, which conforms to the general definition of coefficient. Rho (ρ) adjusts the difference between different correlation coefficients, in other words, the distribution of the output, so that it can become more sparse or compact. From a mathematical point of view, the algorithm measures the sum of the reciprocal of the normalized sub-vectors and the L1-norm distance of each dimension of the parent vector, and maps it to the interval of 0~1, as a strategy to measure the relevance between the parent-child vectors.

The gray coefficients of Shanghai Stock Index data is shown in the table:

Volume	ma20	ma10	ma5	Open	Low	High
0.803750	0.854522	0.860737	0.864776	0.993288	0.993288	0.995271

Table 1. Gray Coefficients for Shanghai Stock Index

Both GRA and PCA conduct dimensionality reduction for the number of eigenvalues. Because the obtained data samples remove invalid columns, there are still 5 features left. In order to make full use of the original data, the number of features finally put into use is selected as 5.

4.5 Grid Search for Optimal Parameters

Grid Search: an exhaustive search parameter modulation method, which finds the best performing parameter by iterating through all the candidate parameters and trying every possibility. It's like looking for a maximum in an array. Why grid search? Take a model with two parameters as an example. There are three possibilities for parameter A and four possibilities for parameter B. All the possibilities are listed and can be represented as a 3*4 table, where each cell is a grid. The best parameter is then represented as a point in the gird, which intuitively is the point with the highest score.

4.6 Model Selection

The stock data set, in which the dependent variable is represented by rise and fall, is a binary classification problem. SVM(Support vector Machine) was originally developed to solve dichotomy problems, and the advantages of SVM are as follows:

1) Support vector is the training result of SVM, which plays a decisive role in the SVM classification decision making process;

2) The final decision function of SVM is only determined by a small number of support vectors, and the complexity of calculation depends on the number of support vectors rather than the dimension of the sample space, which in some way avoids the "dimension disaster";

3) Support vector machine is a convex optimization problem, which can avoid the problem of local minimum.

The SVM algorithm is implemented through the Sklearn library, where the optimal values for parameters C and Gamma can be attained using a grid search. The data set is divided by the ratio of 7:3, and then the prediction process is repeated, or so called periodic: The following day was predicted with the top 70 percent of the data, and the next prediction was augmented with the previous day's results along with the original training data. Therefore, through repeated training and prediction, we can get the final prediction result.

4.7 Results

4.7.1 Results for Optimal Parameters

Optimal value for Gamma (g) value used in SVM is obtained through Grid Search, and represented in graphics below:

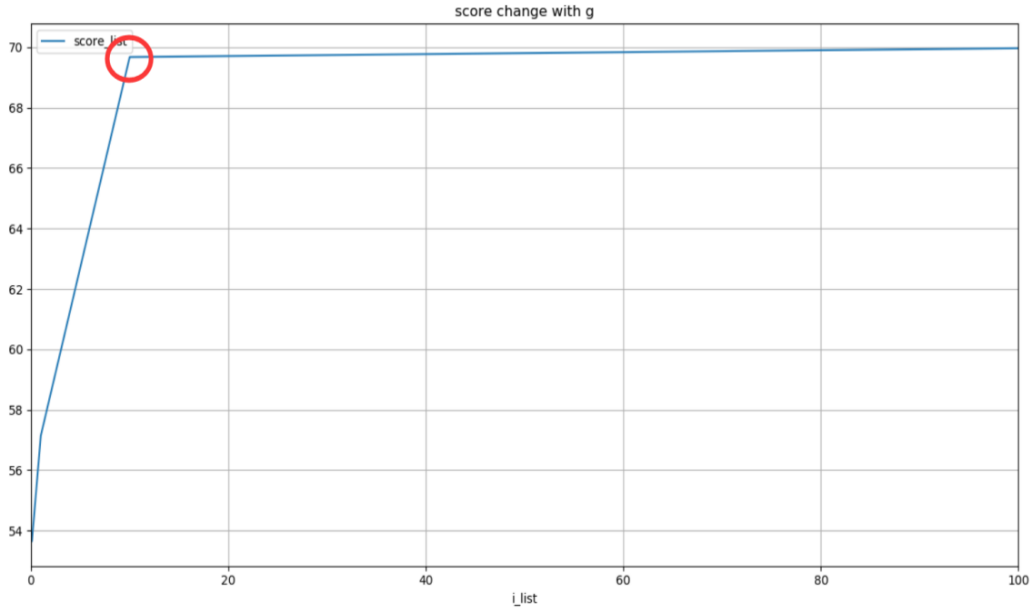

Figure 4. Optimal Gamma Value (g) attained through Grid Search

According to the results, we choose gamma = 10 for best performance.

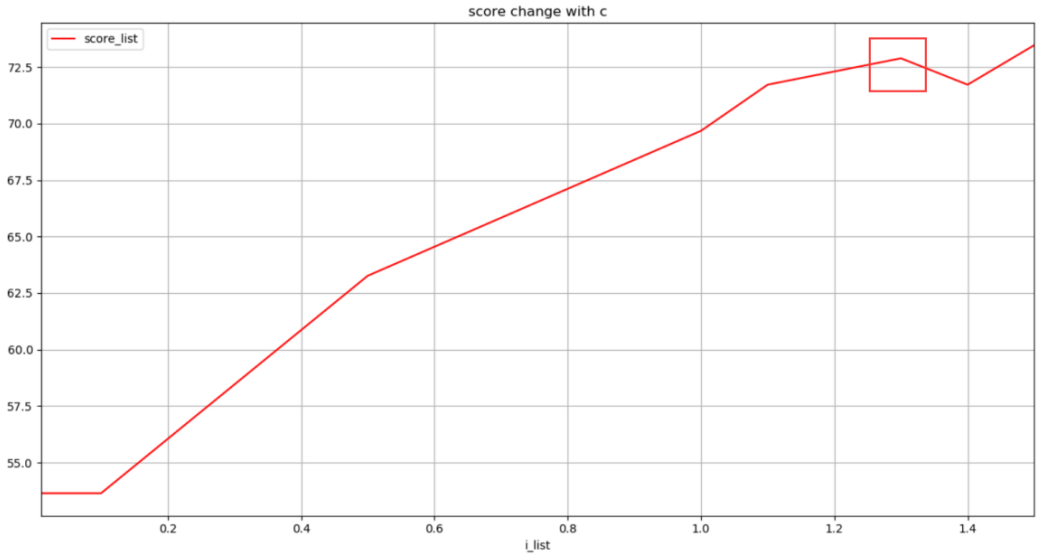

Figure 5. Optimal C Value attained through Grid Search

According to the results we chose C = 1.3 for best performance.

4.7.2 **Results for Accuracy**

We set up the accuracy score as the percentages of the ones that are successfully (correctly) predicted. As we run the experiment, the result shows that, the success rate of the PCA method with SSE stock turns out to be higher with a value of 78.13%, while that of the GRA method only reaches 73.47%. In this case, we test the Baidu stock through the model parameters trained by the Shanghai Composite Index and PCA. The success rate reaches 63.20% under PCA, and 58.71% for the GRA approach.

4.7.3 Results for Assessment and Comparisons

To asses the model, we graph the ROC curve and compute the corresponding AUC for each method.

ROC Curve is short for Receiver Operating Characteristic Curve. The abscissa of ROC curve is False Positive Rate (FPR), and the ordinate is True Positive Rate (TPR), where the calculation methods of FPR and TPR are as follows:

$$FPR = \frac{FP}{FP + TN}$$

$$TPR = \frac{TP}{TP + FN}$$

Figure 6. Formulas for TPR and FPR

FP represents the number of samples that are actually negative but predicted to be positive, TN represents the number of samples that are actually negative but predicted to be negative, TP represents the number of samples that are actually predicted to be positive, and FN represents the number of samples that are actually positive but predicted to be negative.

AUC (Area Under Curve) is the Area Under THE ROC Curve, which can quantitatively reflect the model performance measured based on the ROC Curve.

The graphs are presented below with their respective AUC values:

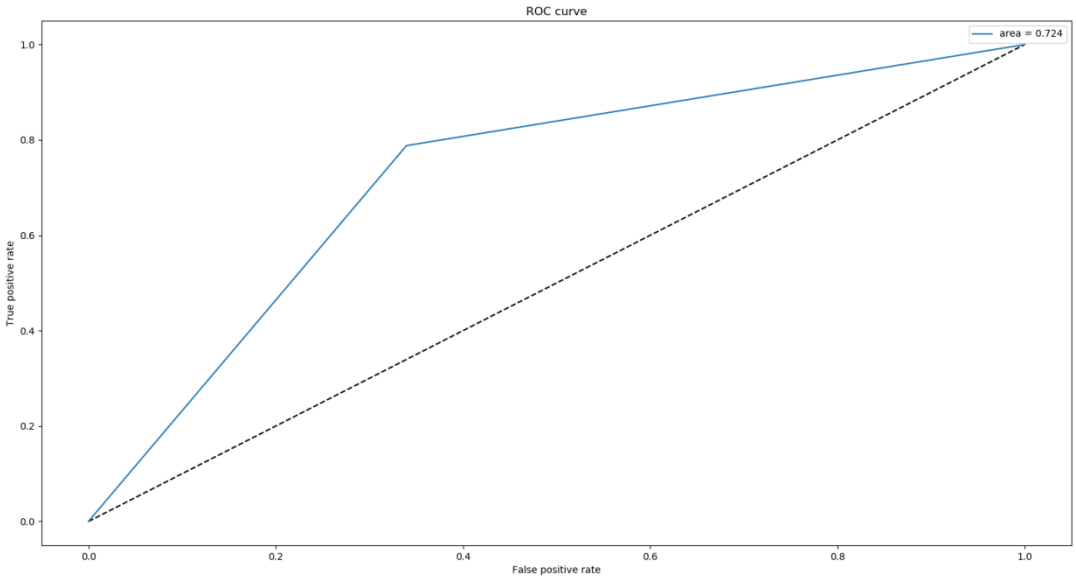

Figure 7. ROC Curve for SSE under GRA

ROC curve of SSE index data as shown in the figure using GRA with the number of features set to be 5. The AUC is 0.7242104183757179.

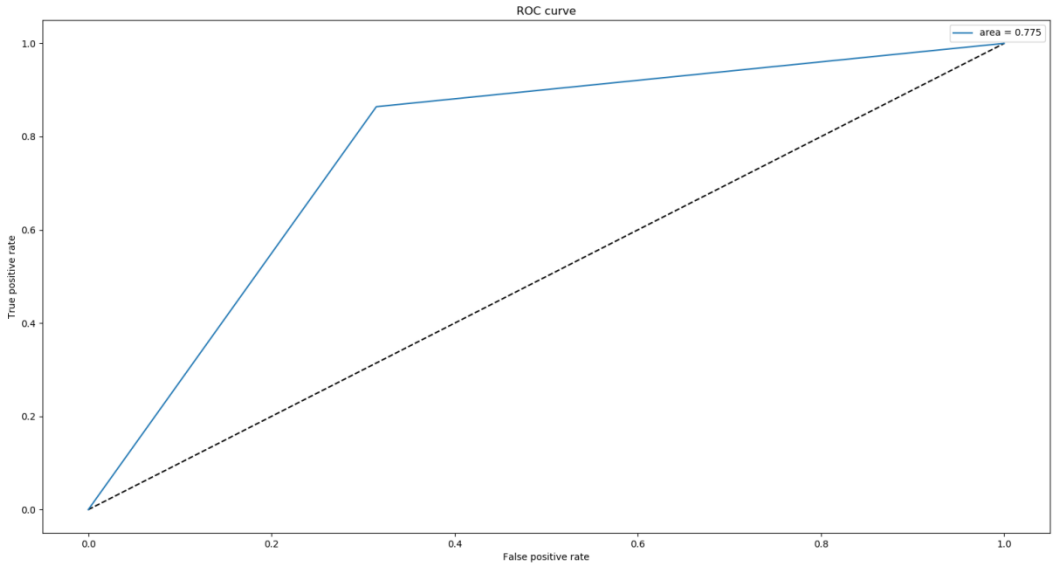

Figure 8. ROC Curve for SSE under PCA

ROC curve of SSE index data as shown in the figure using PCA with the number of components set to be 5. The AUC is 0.7748325129887887.

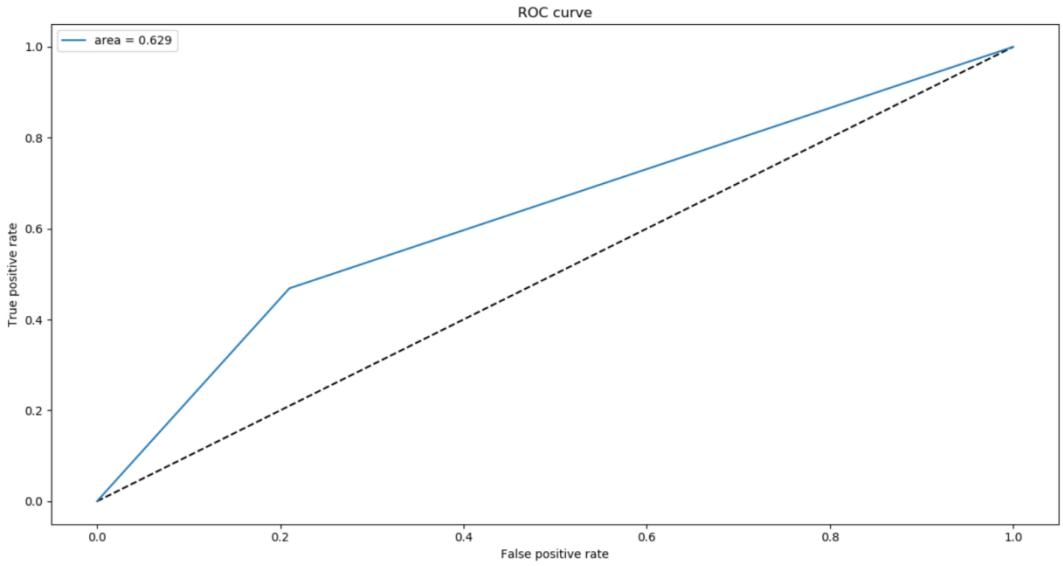

Figure 9. ROC Curve for Baidu under PCA

ROC curve of Baidu stock data as shown in the figure using PCA with the number of components set to be 5. The AUC is 0.6293133385951065.

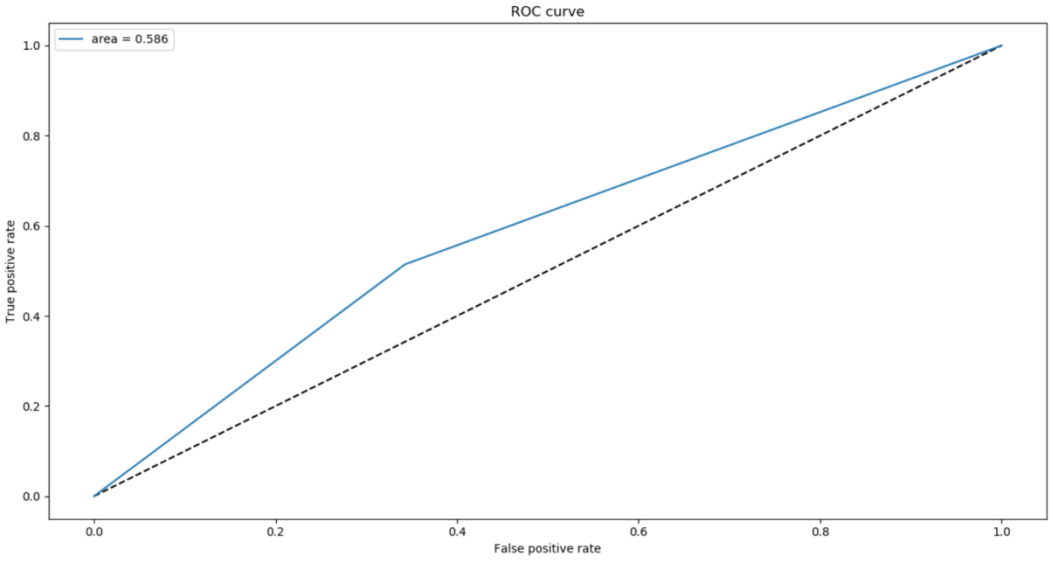

Figure 10. ROC Curve for Baidu under GRA

ROC curve of Baidu stock data as shown in the figure using GRA with the number of features set to be 5. The AUC is 0.5858721389108129.

To have a clearer view, we create a table presented as below.

	PCA(n_components=5)	GRA(feature=5)
SSE stock data	AUC=0.774 Accuracy:78%	AUC=0.72 Accuracy:73.47%
Baidu stock data	AUC=0.629 Accuracy:63.20%	AUC=0.585 Accuracy:58.71%

As shown in the table, generally speaking, PCA has a better performance than GRA in data processing for the sample data we selected.

5. Discussion

The overall performance and the accuracy of the model is in an acceptable range, although the model could be better. The daily stock goes up or down is affected by many factors, such as the revenue of a company, news, policies, etc., and these data are difficult to obtain. Thus, this model should be regarded as a guide and a first step forecaster to predict the direction of the stock price. To make a valid prediction of the value of stock prices, we still need to consider current affairs and acquire more information of the companies we wish to study.

6. Contribution

Chao Zhou: 1) Data Collection and pre-processing
2) GRA approach and parameters tuning
3) Results analysis and comparison
4) Project presentation and Report writing

Yuan Wang: 1) Data Collection and pre-processing
2) PCA approach and parameters tuning
3) Performance assessment and discussion
4) Report writing

7. Code:

The dataset and the source code can be found in the following link: https://github.com/Ro0k1e/CS334_Final_Project