Hong Kong Stock Index Forecasting

Tong Fu tfu1@stanford.edu
Shuo Chen cslcb@stanford.edu
Chuanqi Wei chuanqi@stanford.edu

***Abstract* — Prediction of the movement of stock market is a long-time attractive topic to researchers from different fields. The Hang Seng Index (abbreviated: HSI) is a free float-adjusted market capitalization-weighted stock market index in Hong Kong. We believe the HSI is influenced by other major financial indexes across the world and try different machine learning method to get a prediction based on the history data. The higher accuracy rate we can get, the more confidence we will develop a profitable trading strategy. After several trials of optimization in features and models, we get a 64.96% accuracy rate of the future market trend of HSI.**

1. Introduction

The Hang Seng Index (abbreviated: HSI) is a free float-adjusted market capitalization-weighted stock market index in Hong Kong. It is used to record and monitor daily changes of the largest companies of the Hong Kong stock market and is the main indicator of the overall market performance in Hong Kong. These 48 constituent companies represent about 60% of capitalization of the Hong Kong Stock Exchange. As HK is one of the financial centers across the world, the HSI is considered as one of the most important global financial indicator.

In this project, we plan to find some driven factors behind the market movement that can affect HK market and then use the machine learning methods we learned in class for future market trend prediction. If our learning methods lead to a fairly accurate estimate of the next day's index trend based on the information we have today, we can develop profitable trading strategies based on our predictions.

Our report is organized as follows, section 2 will be data collection and data processing, section 3 will be our basic machine learning models based on our data and an initial prediction of the market trend. The initial prediction result is not so satisfactory so we will move on to optimize our model. In section 4 we fixed the SVM model and try different kernels and parameters for SVM improvement, in section 5 we use feature selection to reduce variance of our test result. Section 6 will present the results and analysis of the whole models.

2. Data Processing

We believe that the performance of other major indexes will have an impact on HSI since the global market is closely connected. We also want to consider major currency rates and commodity price as our potential features. All the features we selected are listed in the Appendix.

We then give an initial correlation analysis based on some of the most significant features.

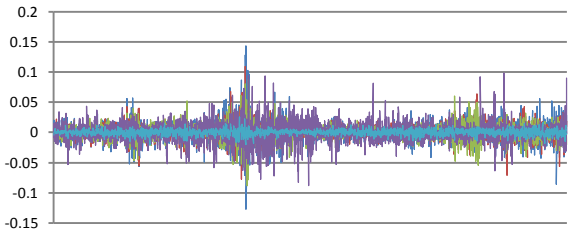

Figure 1: Major Feature Correlation Analysis

We collected the data of 1/1/2000 to 12/31/2012. What we really want is not the absolute value of those variables so we calculate the percentage change of the data as our predictors. We then match our predictors with our response which is the HSI index. What we should be careful here is when we predict the HSI of certain day, all the information we know is up to the day before, so we will match our response with the previous trading day's predictors. This is very important since we later try regression based on the same day's information and the result will be totally changed.

Another consideration is the holidays within different markets. Since our predictors cover different financial products in different regions, data will be missing within different time periods. We will delete the dates when any of the predictors or response is missing.

Finally, we will change the percentage change in HSI into {0, 1} for the market trend and we successfully finished our data processing for the classification problem.

3. Methodologies

3.1. Model Selection

We employed supervised learning models learned in class, including Logistic Regression, GDA, SVM and Naïve Bayes in forecasting HSI.

To test model performance, we conducted cross validation and randomly divided the given data into 70% for training and 30% for testing.

Initially, we trained these four basic models with all features collected and then verified their accuracy using testing dataset, in order to draw a preliminary conclusion on how well each type of model work in predicting HSI trend. Cross validation is implemented here and the following table shows test results of the four models.

Model	Logistic	GDA	NB	SVM
Accuracy	56.13%	55.04%	51.98%	54.18%

Table 1: Accuracy rate for 4 basic models after initial training and testing

From the table above, we can conclude that Logistic regression, GDA and SVM performed significantly better than Naïve Bayes model, whose accuracy rate is as low as 51.98%. Considering that in theory, random guess will lead to an accuracy rate close to 50%, NB model does not make too much sense here due to low accuracy. Therefore, we selected the other three basic models to further improve prediction accuracy but dropped Naïve Bayes.

The reason why Naïve Bayes classifier failed to perform well here is most possibly that Naïve Bayes assumption does not hold in this case, where y (Direction of HSI) is correlated with features but not conditionally independent.

Next, we focused on improving Logistic, GDA and SVM models by feature and parameter selections.

3.2. SVM Improvements

SVM has shown its relatively good performance in predicting HSI compared to the other three basic models. And the following four steps are gone through to improve SVM classifier.

3.2.1. Data Scaling

Scaling data before applying SVM models is significant in avoiding the scenario that features with larger numeric range in value dominate those with small numeric range. Additionally, this helps to reduce computation workload.

In our project, we scalded feature values linearly into a range of [-1, 1]. After being implemented into basic SVM model (with linear kernel), its accuracy rate has increased by 3.29%, as shown in the table below.

Mode	Before Scaling	After Scaling
Accuracy	54.18%	57.47%

Table 2: Accuracy rate before and after data scaling for SVM (linear)

3.2.2. Kernel Selection

We have tried three different kernels in SVM. They are linear kernel, polynomial kernel and Gaussian kernel (RBF, radical basis function) respectively.

The accuracy results of SVM using the three kernels (after scaling) are shown in the table below.

Kernel	Linear	RBF	Polynomial
Accuracy	57.47%	62.38%	59.18%

Table 3: Accuracy rate of different kernels

RBF Kernel achieved the highest accuracy in predicting. This can be attributed to RBF kernel's function to map samples non-linearly into a higher dimensional space and that it has fewer hyper-parameters than polynomial kernel which has an impact on the complexity of model selection.

3.2.3. Feature Selection

We used backward search to select features for different models. The reason of choosing backward search as a feature selection algorithm and procedures of implementation are described in detail in '3.3. Feature Selection' Section.

Mode	Before	After
Accuracy	62.38%	64.32%

Table 4: Accuracy rate before and after feature selection

As we can see, the accuracy rate increased after feature selection.

3.2.4. Parameter Selection

Two parameters in RBF kernel needs to be optimized, namely C and γ. Later on, we employed parameter search to determine the values, with the objective to find a pair of C and γ which works to achieve higher accuracy rate.

We used cross-validation here by dividing the training set into n subsets of equal size, and sequentially, one subset was tested using the classifier trained on the rest of n-1 subsets. Grid-search method was used on C and γ by cross-validation, and the pair which produced the best cross-validation accuracy is selected. Eventually, we set C = 28 and γ = 0.0064182.

After parameter selection, our prediction accuracy rate has been improved further as shown in the table below.

Mode	Before	After
Accuracy	64.32%	64.96%

Table 5: Accuracy rate before and after parameter selection

3.3. Feature Selection

In the previous section, the set of 13 features that we assume could affect the HSI movement was used to train supervised models. We suspected that the number of features is so large that some features might provide overlapping information. Under this situation, some features are redundant given the other features in the selected feature set. To deal with the potential over-fitting, a feature selection algorithm is necessary in reducing data complexity and improving prediction accuracy.

In general, feature selection algorithms designed with different evaluation standard fall into two categories: the filter methods and the wrapper methods. While filter methods acquires no feedback from classifier, the wrapper methods are classifier-dependent, which works for the prediction of HSI movement. An appropriate wrapper method can be used to evaluate the "goodness" of the selected feature subset directly and yield better performance. Meanwhile, the high computational complexity of wrapper methods can be balanced by sample size, which further substantiates the use of wrapper methods in feature selection.

As it is impossible to train model for all possible feature subsets, we require feature selection algorithms to proceed greedily. Forward selection method and backward selection method can be regarded as a possible algorithm to select feature subset yielding best accuracy rate. However, the HSI movement can't be driven by a single factor itself. It is difficult for forward to select a set of features that can work with other features to give the prediction of the direction of HSI movement. The first step of forward search seems to be random because none of features listed above is a good predictor by itself. Therefore, we implement backward search algorithm in the part of feature selection to reduce over-fitting problem. During the selection, we calculate the estimated accuracy rate of the learning algorithm for each set of features by computing the generalization error using hold-out cross validation. In this way, we

can train learning models with a set of selected features and the corresponding accuracy rate is summarized in the table below.

Model	*Logistic*	*GDA*	*NB*
Before	56.13%	55.04%	51.98%
After	60.02%	56.88%	53.04%
Model	*SVM(Linear)*	*SVM(RBF)*	*SVM(Poly)*
Before	57.47%	62.38%	59.18%
After	58.24%	64.32%	61.87%

Table 6: Accuracy rate before and after feature selection (compute without implementing parameter selection for SVM)

Logis-tic	**SVM-Linear**	**SVM-RBF**	**SVM-Poly**	**NB**	**GDA**
FTSE	*S&P*	*S&P*	*S&P*	*S&P*	*DJIA*
SSE	*FTSE*	*Nikkei*	*Nikkei*	*FTSE*	*FTSE*
STI	*CNY/$*	*FTSE*	*FTSE*	*CNY/$*	*SSE*
NASD	*Euro/$*	*SSE*	*SSE*	*STI*	*CNY/$*
		JPY/$	*JPY/$*	*Gold*	*NASD*
		Gold	*Gold*		*Gold*

Table 7: Features selected for different models

4. Results and Analysis

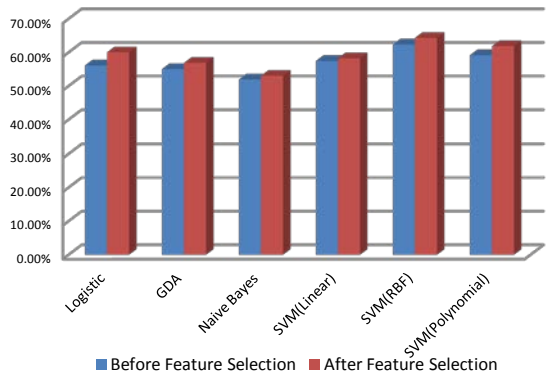

Figure 2: Accuracy rate before and after feature selection

At the beginning, the initial correlation analysis indicates the hidden interconnection between the Hang Seng index and global markets indices that close right before the Hong Kong trading time. To unmask this hidden relationship, we proposed several machine learning based models for predicting daily trend of Hang Seng index. Each of the forecasting models described in the section 3 was estimated and improved through feature selection. Specifically, the actual model performance was evaluated by holdout cross validation. We were trying to test whether specific machine learning model trained on market data can be applied to make predictions about future market trends. At this stage, the relative performance of the models was measured by empirical error rate when predicting the direction of HSI movement on the hold out cross validation set.

The performances of four types of supervised learning models are shown in the figure above. Each model is able to give a prediction rate higher than 50%, which is better than random. Those raining models can absorb the relevant information of other financial markets on the previous trading day in order to predict the trend of HSI movement. Among all learning models, according to the result table, SVM with Gaussian Kernel gives the best performance in predicting index movement, whose accuracy rate is 62.38%. After further improvement by backward search algorithm and parameter selection, the rate of accuracy goes up to 64.96%.

Besides high accuracy rate, SVM is the most efficient tool in countering the over-fitting problem. The features we included in learning at first are high-dimensional, which brings over-fitting problems to these learning models. Comparing the prediction rate with and without the feature selection, we can figure out that certain feature selection algorithm plays a more significant role in reducing the over-fitting problem in SVM than in other classification models. Therefore, SVM has proven to a possible solution in prevent over-fitting problems given limited training dataset.

5. Conclusion and Furtherwork

Predicting stock market movements has always been a challenging task considering the trends being

affected by various random factors such as volatility or other macroeconomics factors. However, the interaction among all financial indices allows us to capture those changes in order to make profit in stock trading. In this project, we investigated the use of several machine learning models to predict HSI movement direction. Even if the learning models we used in this project are just simple derivation of supervised learning we learned in the class, they were able to meet our expectation of achieving modest accuracy rates and capturing the relevant features for each training model. Although logistic regression generates a decent prediction result, of which the accuracy rate is over 50%, SVM is considered to be a promising type of tool for HSI movement forecasting. This is quite reasonable considering the way we select features for forecasting models. The initial screening of features is based on macroeconomic analysis. It is evitable that over-fitting problem will arise because of the high-dimensionality of features. As demonstrated in both empirical analysis and actual training results, SVM has been shown to be very resistant to the over-fitting problem, eventually reaching a high generalization performance and improving profitability and the stability of predictions. The result after feature selection further confirms that SVM is superior to the other individual classification methods in forecasting daily movement direction of Hang Seng Index, because it performs best among all the forecasting models. This can be interpreted as an indication for financial analysts and traders, which can bring a certain level of capital gain.

However, each method has its own strengths and weaknesses. Also, there are several aspects of our research that need to be improved in the future. For example, we assume different features can provide relevant information for HSI movement forecasting over different periods of time. Including several time series features can greatly help to detect the underlying relationship between HSI movement and the distribution of financial time series. After all, our experience in this project has shown the great potential that machine learning has in the field of financial forecasting. We believe exploration in improving performance via better feature selection algorithms as well as optimizations can generate a more convincing prediction result.

6. References

[1] Y.S. Abu-Mostafa and A.F. Atiya “Introduction to financial forecasting,” Applied Intelligence, 6 (1996), pp. 205-213.

[2] R. Choudhry and K. Garg, "A hybrid machine learning system for stock market forecasting," Proceedings of World Academy of Science, Engineering and Technology, vol.29, pp. 315-318, 2008.

[3] W. Huang, Y. Nakamori and S. Wang “Forecasting stock market movement direction with support vector machine,” Computers & Operations Research, 32, pp. 2513–2522005, 2005.

Appendix

1. Feature Table

Feature	Category	Explanation
S&P 500	Stock Index	Standard & Poor's 500, a stock market index based on the market capitalizations of 500 large companies having common stock listed on the NYSE or NASDAQ
DJIA	Stock Index	Dow Jones Industrial Average
Nikkei	Stock Index	Nikkei Stock Average, a stock market index for the Tokyo Stock Exchange
FTSE	Stock Index	FTSE 100, a share index of the 100 companies listed on the London Stock Exchange with the highest market capitalization
SSE	Stock Index	SSE Composite Index, a stock market index of all stocks (A shares and B shares) that are traded at the Shanghai Stock Exchange
Crude Oil	Commodity	Commodity price of crude oil
CNYUSD	Currency Rate	Chinese Renminbi-US Dollar exchange rate
JPYUSD	Currency Rate	Japanese Yen-US Dollar exchange rate
EuroUSD	Currency Rate	Euro-US Dollar exchange rate
AUDUSD	Currency Rate	Australian Dollar-US Dollar exchange rate
STI	Stock Index	Straits Times Index, it tracks the performance of the top 30 companies listed on the Singapore Exchange
NASDAQ	Stock Index	A stock market index of the common stocks and similar securities (e.g. ADRs, tracking stocks, limited partnership interests) listed on the NASDAQ stock market
Gold price	Commodity	Commodity price of gold