Birdie Agent

MSBD 5013 Project Final Report

Gan Weisen 20446377
Luo Wenting 20461341
Li Yunzhu 20446391
Chen Weizhu 20450990

BACKGROUND

Future investment is increasingly becoming an indispensable part of people's life, but the return of future is often proportional to the risk, that is to say the higher the return, the greater the risk and vice versa. An effective method of analysis, especially one that maximizes profitability and reduces risk, is highly desirable. However, since the future price can be regarded as time series data, many random factors can have impact on the future price more or less. Therefore, the future price shows complex nonlinearity and uncertainty. It is hard to reveal its inherent laws using traditional time series prediction techniques. Our project is to analyze the price data of futures, build a reasonable statistical prediction model and put forward a trading strategy in order to achieve the stability of high returns.

1. EXPLORARY DATA ANALYSIS

1.1 MODEL ATTEMPT

The price of future can be regarded as a random time series data and there are many methods, such as Grey system, Chaotic time series prediction and Neural Network, to do time series prediction. As for Grey system, it views the behavior of the system as the process of random change and it is consistent with the characteristics of futures to some extent. This model is effective when finding the statistical law from a large amount of historical data, however, the data of futures we get is limited and when facing with a poor information system with a small amount of data, the analysis is more difficult. A neural network has a better result in dealing with a non-linear system, but there are two major defects in the convergence process of a conventional BP nerve network, namely the slow convergence rate and the problem of local minimum. Also with the improvement of training ability, the ability to predict will decrease, if we cannot balance the relationship of them, the prediction effect will not be good. The method of SVM have systematically studied some fundamental problems in pattern recognition with limited samples, which largely solved the problem of model selection and over-fitting in our project. Not only can SVM separate two classes, but also make classification interval be maximum. At the same time, the most important feature of the economic system is non-linear. The most crucial problem that support vector machines can apply in the economic field is that they can handle nonlinear problems by transforming into a high-dimensional space and find an optimal hyper-plane, it is well-suited to predict whether the price of futures is up or down. So in the first week, we tried the model of SVM.

1.2 FEATURE EXTRACTION

The given historical data of future includes the opening price, closing price, the lowest price and the highest price, but all these data are the price. However, the price does not express the characteristics of future well, that is to say, the price of futures is not likely to rise or fall to a large extent after reaching a price range. So we extract 5 features: the opening price, closing price, the lowest price, the highest price and daily closing price change. Based on the futures' historical price, we treated the model as a two class classifier. If the price of future is up, we labeled it as '1'. If the price of future is down, we label it as '0'. Based on this classification rule, we used historical data as training data.

In order to have a better performance, we adjusted parameters of SVM model and also changed the kernel function. However, we found that the model with default parameters and Radial Basis Function kernel function has the highest prediction accuracy.

Kernel Function	Linear	Polynomial	RBF	Sigmoid
Cross Validation (Accuracy)	32.54%	55.62%	67.98%	57.47%

Figure 1 Accuracy of different kernel function (5 fold)

1.3 MODEL CHANGE AND MODEL SELECTION

In the later exploration, we think that only using SVM will limit our ability to predict the trend of futures, so we try to use ensemble method to improve the prediction accuracy. The ensemble methods improve the accuracy of classification by aggregating the predictions of multiple classifiers. We attempt to use typical ensemble methods (random forest & Adaboost) to achieve high accuracy.

In the algorithm of Adaboost, we also want to classify the futures. The general idea is to establish a series of classifiers. The weight of each sample in the next training is determined the result of each weak classifier. If the classification is correct, the weight will decrease; if the classification is wrong the weight will increase. Finally each weak classifier combines into a strong classifier.

The Optimal Decision Tree algorithm is based on heuristic algorithms, such as Greedy Algorithm, which is based on local optimal solution. Heuristic algorithm cannot guarantee the return of the global optimal Decision Tree. Decision Tree algorithm may produce too complicated trees, it is not good for getting a common model of data. It may lead to over-fitting problem. The Random Forest algorithm can solve the over-fitting problem at most of time. However, Random forests have been proved to be over-fitting when applied to some situation that may contain too much noisy. For the data with different values of attributes, the attributes with too much values are more likely to have a greater impact on the random forests algorithm. Therefore, the attribute weights generated by Random Forest Algorithm on such data cannot be trusted.

The results of these two models can be explained in the later part. As the result is not that satisfactory, we finally decide use SVM to train our model.

2. STATISTICAL MODEL

The previous weekly models are trained for a short period of time, so we hope to pick a more universal way to adapt to different situations. So we use a long period of data to train our model and use variable-controlling approach to do the experiement--we keep data and strategy unchanged and only change our statistical model).The following backtest results shows that the three model (SVM, random forest and Adaboost) each has its own characteristics.

2.1 SVM

	Backtest
Annual return	1.4%
Cumulative returns	0.6%
Annual volatility	11.6%
Sharpe ratio	0.18
Calmar ratio	0.26
Stability	0.00
Max drawdown	-5.5%
Omega ratio	1.04
Sortino ratio	0.27
Skew	0.51
Kurtosis	3.83
Tail ratio	0.99
Daily value at risk	-1.4%
Alpha	0.00
Beta	1.00

Worst drawdown periods	Net drawdown in %	Peak date	Valley date	Recovery date	Duration
0	5.48	2017-09-01	2017-10-30	2017-11-22	59
1	5.32	2017-12-01	2017-12-07	NaT	NaN
2	0.95	2017-11-24	2017-11-28	2017-11-29	4
3	0.72	2017-08-24	2017-08-29	2017-09-01	7
4	0.42	2017-08-07	2017-08-14	2017-08-16	8

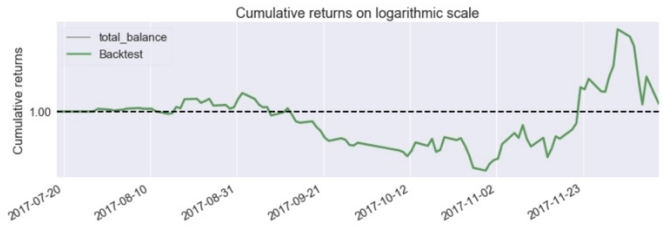
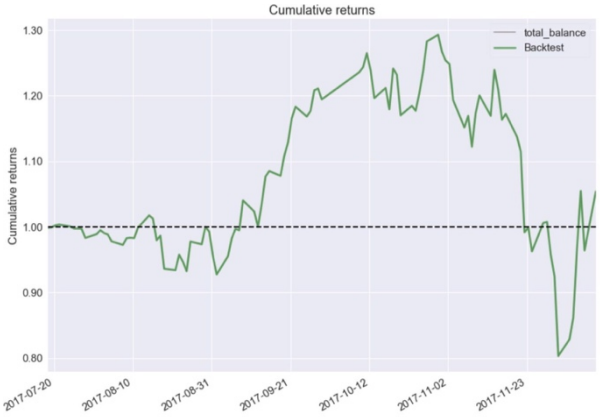


Figure 2 Backtest of SVM

2.2 RANDOM FOREST

	Backtest
Annual return	14.1%
Cumulative returns	5.4%
Annual volatility	57.3%
Sharpe ratio	0.52
Calmar ratio	0.37
Stability	0.10
Max drawdown	-37.9%
Omega ratio	1.10
Sortino ratio	0.74
Skew	-0.22
Kurtosis	2.84
Tail ratio	0.99
Daily value at risk	-7.1%
Alpha	0.00
Beta	1.00

Worst drawdown periods	Net drawdown in %	Peak date	Valley date	Recovery date	Duration
0	37.86	2017-10-30	2017-12-01	NaT	NaN
1	8.83	2017-08-14	2017-09-01	2017-09-08	20
2	7.48	2017-10-11	2017-10-20	2017-10-27	13
3	3.86	2017-09-08	2017-09-12	2017-09-14	5
4	3.08	2017-07-21	2017-08-07	2017-08-14	17



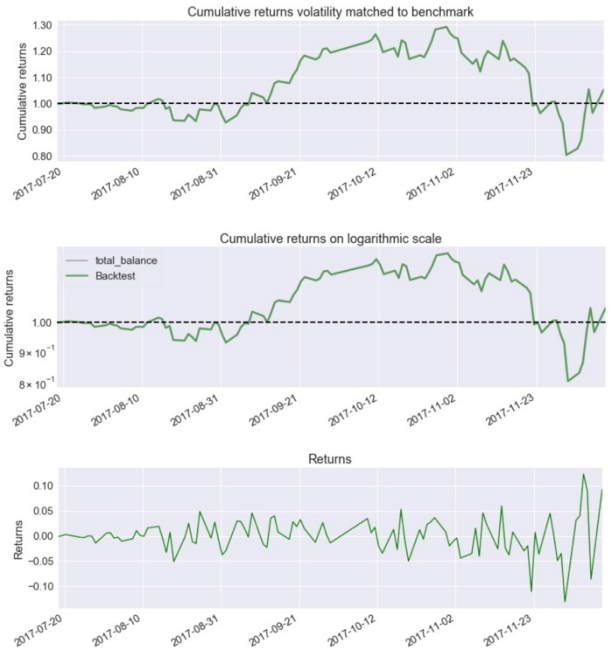


Figure 3 Backtest of random forest

2.3 ADABOOST

	Backtest
Annual return	2.7%
Cumulative returns	1.1%
Annual volatility	69.8%
Sharpe ratio	0.34
Calmar ratio	0.09
Stability	0.38
Max drawdown	-28.4%
Omega ratio	1.15
Sortino ratio	0.76
Skew	6.24
Kurtosis	59.1
Tail ratio	1.02
Daily value at risk	-8.7%
Alpha	0.00
Beta	1.00

Worst drawdown periods	Net drawdown in %	Peak date	Valley date	Recovery date	Duration
0	28.44	2017-09-01	2017-12-07	NaT	NaN
1	1.76	2017-08-10	2017-08-15	2017-08-16	5
2	1.45	2017-08-28	2017-08-30	2017-09-01	5
3	0.61	2017-08-18	2017-08-25	2017-08-28	7
4	0.42	2017-08-16	2017-08-17	2017-08-18	3

Figure 4 Backtest of Adaboost

ANALYSIS:

Annual volatility of SVM is 11.6%, which is much lower than that of Random Forest and Adaboost. It means that the return of using SVM will be much more stable and when the market is volatile, using SVM we will not suffer much. We can see it from Max drawdown, it is -5.5%, -28.4% and -37.9% respectively for using SVM, Adaboost and Random forest. Daily value at risk is -1.4%, -8.7% and -7.1% respectively for using SVM, Adaboost and Random forest.

The Calmar ratio is a comparison of the average annual compounded rate of return and the maximum drawdown risk of commodity trading advisors and hedge funds. The lower the Calmar Ratio, the worse the investment performed on a risk-adjusted basis over the specified time period; the higher the Calmar Ratio, the better it performed. The Calmar Ratio of using Random Forest is the largest. The Calmar Ratio of using Adaboost and SVM are acceptable.

On the whole (we can see from the Figure 4), using SVM is low risk and low return. Using Random forest is high risk and high return. However, using Adaboost is high return and extremely high risk. In this point of view, we can eliminate Adaboost for it is not out-performed when compared with the other two methods. And from the content we discussed before, we need a strategy which is not that aggressive, so here using SVM is the best choice.

	risk	return
SVM	low	low
Random forest	high	high
Adaboost	extremely high	high

Figure 5 Comparison of three methods

3. TRADING STRATEGY

Based on the prediction of our model, we put forward our strategy. At first our strategy is intuitive and clear. We trained our model by using the data of 13 futures. And we selected first few futures that can make our weekly return maximum. For each selected future, if we predict it will go up (that is to say the label = 1), and the cash we hold is 20000 more than my_cash_balance_lower _limit, then we buy 10 shares. On the other hand, if we predict the future we selected will go down (that is to say the label = 0), and the cash we hold is judged 40000 less than my_cash_balance_lower_limit, then we sell 10 shares.

In the process of our exploration, we changed the training model, and we adjusted the strategy at the same time. According to the weekly return of last week's strategy, we modified our new strategy—if the weekly return of last week's strategy is positive, we tried another model and kept last week's strategy; if the weekly return of last week's strategy is negative, we changed our strategy. The first strategy we changed is that we sorted 13 futures from largest to smallest according to its daily return. At first we bought 30 shares of top 3 futures and transact every day. Every time we transact, if the futures in top 3 are held, we will continue to held these futures. If not held, then we sold the futures that are not in the top 3 and buy the futures that in the top 3. The second strategy we changed is that if we predict that the price of futures will be 50% higher than today's price, then we buy these futures and each 30 shares; if we predict that the price of futures will be 30% lower than today's price, we sell all these futures. (Early stop-loss will not suffer much.)

Although it seems that the two strategies we changed in the exploration process are more complex, it is not as effective as the simplest strategy in the first attempt. So

based on the first attempt strategy, we modified parameters and transaction frequency to come up with the final strategy.

In the final strategy, we use SVM as the training model and for each selected future, if we predict its price will go up, and the cash we hold is 100000 more than my_cash_balance_lower _limit, then we buy 10 shares. On the other hand, if we predict the future's price will go down, and the cash we hold is 100000 less than my_cash_balance_lower_limit, then we sell 10 shares. This strategy is little different from the past ones, we transact futures minute by minute, instead of day by day. The transaction frequency we tried is 30 min, 60 min, 120min, 240min.After considering the futures transaction costs and the sensitivity of trading frequency to market volatility, we find that choosing 120 minutes as a trading frequency can achieve good results. And the reason we choose 10 as the transaction shares each time is the more shares we buy, the higher risk we take. Since the future market fluctuates a lot, we hope the return we get is stable, so we chose 10 as transaction shares at each time conservatively.

4. CONCLUSION&FUTURE WORK

There has been a strong interest in applying machine learning techniques to financial problems. In our project, we use the model of SVM to predict the price of futures and come up with a strategy to get as much profits as possible. From the empirical results, the SVM model works fine in the market timing ability to some degree. The model achieved 67.98% correct discrimination in the past 2 months, and the annual return of our investment strategy based on SVM model reach 1.4% and the annual volatility is only 11.6%.Compared with other models, it can achieve low risks and profitable. However, the model also has some shortcomings. The ability of predicting down markets is poor and the transaction signal is more frequent. Although we have tried different models, but the two-month-data is so limited for us to train the best fitting model of the futures market and if we can get more data, we will further improve our model to get a more accurate prediction.

Meanwhile, the selection of the extraction features also has a great influence on the accuracy of the prediction results in the statistical prediction, because we lack of more professional financial knowledge, the feature extraction is not fine defined. If given more time, we would like to fine-extract futures' features to get a further optimize performance.

In addition, the strategy we put forward is more about data rather than logic. Therefore, when the market is changing, it cannot be adjusted in time and effectively. How to effectively apply the machine learning model and theory to the futures market, there is still no good method. The balance point between theory and practice will leave us to explore.

5. ACKNOWLEGEMENT

Our group would like to extend our sincere gratitude to Professor Kani Chen for the guidance in this semester and also we would like to thank our TA for offering us much help during this semester.