Intraday Settlement Risk Forecasting Model Based on SVMs

Yinhua Li

School of Management, Graduate University of Chinese Academy of Sciences

Beijing 100190, China

liyinhua07@mails.gucas.ac.cn, (O) +86-10-8268-0697, (F) +86-10-8268-0698

Yingjie Tian

Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences

Beijing 100190, China

tianyingjie1213@163.com, (O) +86-10-8268-0997, (F) +86-10-8268-0698

Yong Shi

Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences

Beijing 100190, China

and

College of Information Science and Technology, University of Nebraska at Omaha

Omaha, NE 68118, USA

yshi@gucas.ac.cn, (O) +86-10-8268-0684, (F) +86-10-8268-0698

Intraday Settlement Risk Forecasting Model Based on SVMs

Yinhua Li1,2,Yingjie Tian1, Yong Shi1,3[[1]](#footnote-2)\*

1Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences

Beijing 100190, China

2School of Management, Graduate University of Chinese Academy of Sciences

Beijing 100190, China

3College of Information Science and Technology, University of Nebraska at Omaha

Omaha, NE 68118, USA

Abstract

In order to predict the intraday liquidity risk of settlement members in financial futures market, we first transform the prediction problem to be a classification problem, and then apply the standard support vector machine (SVM) to construct models in advance 5 minutes, 30 minutes, 55 minutes, 80 minutes and 105 minutes respectively. Two feature selection methods based on F-score and forward-backward stepwise are used to combine with SVM. Numerical experiments show that our prediction models are stable and efficient. The forward-backward stepwise method can select more important predictors, while the model in advance 5 minutes has more prediction power.

Key words

Feature selection; Settlement risk; SVMs; F-score; Stepwise

## 1. Introduction

With increasing availability of the high frequency data, the task of risk management has started to focus on the intraday data level gradually. At the level of application, models of high-frequency data provide a quantitative framework for market making and optimal executive of trade. Moreover, the study of high frequency market dynamics is important for risk management and regulation[[1](#_ENREF_1)]. There are several research focuses. The literature branches in two broad directions. One of the most popular themes is forecasting intraday volatility of stock index futures price[[2](#_ENREF_2), [3](#_ENREF_3)]. Another theme is how to set margin level in order to avoid fund liquidity risk [[4-7](#_ENREF_4)].

In order to analysis the intraday volatility or model the price change by stock index futures product, one of the most broadly used approach is to considered stochastic model using the classical autoregressive moving average (ARMA) or the generalized autoregressive conditional heteroscedastic (GARCH) model. Xie and Li (2010)[[8](#_ENREF_8)] use GARCH(1,1), EGARCH and IGARCH to estimate intraday volatility of S&P 500 stock index futures.

However in China, the margin level is relatively static which is set by China securities regulatory commission. How to build the risk forecasting model and what factors will influence the occurrence of fund liquidity risk are two interesting questions in intraday risk management for China finance futures exchange.

In China, financial future exchange (CFFEX) is the only place where financial futures are traded and settled. According to Article 65 in chapter IV of measures for administration of futures exchange (2007), a futures exchange adopting the graded clearing system only collects margin from its clearing members. Therefore, the risk of all of the clearing members is managed by CFFEX directly. The clearing department is in charge of controlling liquidity risk and credit risk of clearing member. Based on exist measures of risk management, they have motivation to improve risk administration both on philosophy and techniques.

However, machine learning methods have less presented in the research area[[1](#_ENREF_1)]. We will introduce support vector machine model to predict settlement risk for CFFEX.

Classification performance is determined by the inherent class information available in the features vector provided. Depends on these information, we can judge what category the observation is belongs to. The feature selection process is to find a separating plane that discriminates between two point sets in an n-dimensional feature space that utilizes as few of the n features as possible.

This paper will mainly solve two problems. The first problem is how to transfer the settlement risk forecasting problem into the classification problem. The second one is how to select proper features during the model building process. The structure of the paper is as follow. We shall first briefly introduce the settlement risk problem in the second section, then transform the problem of settlement risk forecasting into the classification problem. Based on that, we are able to build risk forecasting model. In the third section and the fourth section, the theory of the standard SVM and two kinds of feature selection methods will be introduced specifically. We will demonstrate whether the SVM suits for solving settlement risk forecasting problem based on realistic data supplied by CFFEX. The sixth section will show the conclusion of the paper.

## 2. Problem Transformation

The margin is classified into the security for clearing and the security for trading. The term “security for clearing” refers to the margin which has not been occupied by any contract. The term “security for trading” refers to the margin which has been occupied by any contract.

Funds utilization is most direct measures to assess the risk level of clearing member. It reflects the ratio of occupied margin in futures trading to total equity. The high value of the funds utilization, the more risk will be exposed to settlement member. Futures trades have the characteristic of high leverage. Therefore, if the market trend moves to the opposite direction of the investors’ expectation, the settlement member will be exposed to the risk of ruin.

The value of funds utilization is continues. According to international conventions and expert experiences, if the value of funds utilization is more than eighty percent, then it indicates that the settlement member is exposed to risk, otherwise the opposite. See equation 2.1,  means funds utilization.

 (2.1)

Through introducing the threshold, we transfer continues variable into a binary-state variable, risk and non-risk samples respectively. If the central settlement house judge a member as risk or not depends on funds utilization, the task of risk management can be transferred into a classification problem. Based on a series of risk related factors, we construct risk model to predict if the funds utilization will exceed eighty percent some periods prior to risk occurrence for the settlement member.

How long will the model predict risk? It requires dividing data into observation period and performance period. In observation period, it requires to survey the behavior of settlement member and produce independent variables x according to the data in this period. After the observation period, the settlement member will perform as risk or non-risk in performance period. The data in the performance period is used to produce the dependent variables y. we collect data in both observation period and performance period, then construct the training data set:

 (2.2)

Where The destination is get a real-valued function in in order to use the decision function

 (2.3)

to judge the performance of the settlement member in performance period based on the data in observation period in future.

Observation period (x)

Performance period (y)

t0

t1

t2

## 3. Standard SVMs

Support vector machines (SVMs), which were introduced by Vapnik and his co-workers in the early 1990's [[9](#_ENREF_9)], are computationally powerful tools for supervised learning and have already successfully applied in a wide variety of fields[[10-15](#_ENREF_10)].

For a binary classification problem with the training set

 (3.1)

where the standard support vector classification (SVC) searches for two parallel hyper planes with maximal width between them, which leads to solving a convex quadratic programming problem (QPP):

 (3.2)

  (3.3)

 (3.4)

where and is a penalty parameter. For this primal problem, solves its Lagrangian dual problem

 (3.5)

 (3.6)

 (3.7)

where is the kernel function, which is also a convex quadratic problem and then construct the decision function.

After getting the solution of the problem (3.5)-(3.7), the of problem (3.2)-(3.4) is calculated as

 , (3.8)

and the decision function with the optimal separating hyperplane can be presented as

 (3.9)

where

 (3.10)


## 4. Feature Selection

Standard SVMs cannot get the important features, while identifying a subset of features which contribute most to classiﬁcation is also an important task in machine learning[[11](#_ENREF_11)]. The advantage of feature selection is twofold. It leads to parsimonious models that are often preferred in many scientiﬁc problems, and it is also crucial for achieving good classiﬁcation accuracy in the presence of redundant features. We can combine SVM with various feature selection strategies, some of them are “ﬁlters”: general feature selection methods independent of SVMs. That is, these methods select important features and then SVMs are applied. On the other hand, some are “wrapper-type” methods: modiﬁcations of SVMs which choose important features as well as conduct training/testing. In this paper, we will apply two ﬁlter methods. They are F-score method and forward-backward stepwise method.

**4.1 F-score**

F-score is a simple and generally quite effective technique. Given the initial training set (3.1), suppose that the number of positive and negative point are and respectively, then the F-score of feature is defined as

 (4.1)

where

 (4.2)

 (4.3)

 (4.4)

The number of indicates the distribution of feature between the positive and negative sets, and the denominator of indicates the one within each of the two sets. The larger the F-score is, the more likely this feature is more discriminative. Therefore, this score can be used as a feature selection criterion.

**4.2 Forward-backward stepwise**

Forward-backward stepwise feature selection is a combination with forward selection and backward selection, in which each forward step is followed by a backward step to remove features in the model that are no longer significantly related to the response, thereby overcoming some of the shortcomings of forward selection[[16](#_ENREF_16)]. With a significance threshold (usual 0.05)[[17](#_ENREF_17)], it starts with no explained feature, then only significant feature enter the model in each step. Meanwhile, the program will examine automatically whether the feature existing in the model becomes non-significant when other feature enter the model. If the Wald statistic[[18](#_ENREF_18)] of the feature exceeds the rejection criteria, it is culled. Otherwise, it will search other features which are significant. The whole procedure terminate until no more features entered nor no one should be dropped.

## 5. Empirical results

In this section, we will apply standard SVC with the above feature selection methods to build the forecasting model based on the data supplied by CFFEX. The database includes high frequency trading and settlement data of 75 settlement members. Each member has 9,460 observations which contain 172 trading day from Apr 16th to Des 31th. The data frequency is 5 minutes. We plan to build risk forecasting models for each member. In numerical experiment, we choose one member who has the largest number of risk observations. The process of the experiment can be described in Fig I.

Fig I The process of the numerical experiment

**5.1 Analysis of the influence factors**

The influence factors of the settlement risk mainly come from two aspects. They are trading behavior of the settlement member and market quotation. The trading behavior is the most directly factors which can influence settlement risk. Five proxy variables are selected to measure the trading behaviors of the settlement member. See Table I. They are trading volume, turnover, net open rate, net position and gross position respectively.

The proxy variable trading volume is an important sentiment index. It reflects the strength of relationship between both sides of supply and demand. One can judge the price trend and price fluctuation intensity by analyzing the relationship between trading volume and price. In general, if the trading volume and the price increase at the same time, it signifies that the price will increase sustained. And if the price increase and the trading volume shrink, it implies the shortage of the investor confidence. Therefore, the price will decrease in high probability.

And another important proxy variable is open interest. The open interest is the total number of outstanding contracts that are held by market participants at the end of the day. The open interest can be classified into buy position and sell position. It is often used to confirm trends and trends reversals for future and options contracts. Similarly, the open interest has close relation with the price of futures. Therefore, the change of the open interest reflects the flow of funds in futures market and it depends on the behavior of the trader. If the open interest increases when the price rises, it implies that new buyers open long position. And the price will rise recently. In the second cases, if the open interest decreases when the price drops, it implies that new sellers open short position. The price will drop. In the third cases, if the open interest decreases when the price rises, it implies that the long position sell their position and leave the market. And in the fourth cases, if the open interest deceases when the price drops, it implies short position buy their position to make a profit. And the price will turn to rise in a short time.

In order to make features more explainable, we introduce three features related to open interest. They are net open rate, net position and gross position. The attributes net open rate is a ratio. The calculation can be seen in Table 1, the fourth row. The larger value the net open rate is, the more risk will be exposed to the settlement member. Another important risk influence factor is market quotation. The price trend of spot and futures are expected factors for the settlement members. The price and yield of the main contract are introduced to signify market quotation.

Table I The explanation and calculation of the features

|  |  |  |  |
| --- | --- | --- | --- |
| Variable name | Variable meaning | Description | Function |
|  | trading volume | Measure the deal statement |  |
|  | turnover | Measure the deal statement |  |
|  | net open rate | open interest |  |
|  | net position | open interest |  |
|  | gross position | open interest |  |
|  | price of the main contract | market quotation |  |
|  | yield of the main contract | market quotation |  |

**5.2 Numerical experiments**

The derivative features always perform more predictable than original features, therefore we design 28 derivative features based on above 7 original features. We calculate the average, variance, range and last time value of the original feature during the observation period mentioned in section 2 as the derivative features.

In order to examine the accuracy of the different forecasting period model, we build five different models for the member, see Table II. They can forecast settlement risk 5 minutes, 30 minutes, 55 minutes, 80 minutes and 105 minutes respectively early before it occurred.

Table II the numerical experiments results

|  |  |  |
| --- | --- | --- |
|  |  | Different observation period  |
|  |  | 5min | 30min | 55min | 80min | 105min |
| No selection | C | 26.85 | 39.13 | 39.13 | 22.644 | 348.89 |
|  | Overall Accuracy | 0.9825 | 0.9782 | 0.9651 | 0.9782 | 0.9869 |
| F-score | features | Var16,13,20,17,9 | Var16,13,20,17,9 | Var16,13,20,9,17 | Var16,13,20,17,9 | Var16,13,20,17,9 |
|  | C | 348.88 | 348.88 | 348.89 | 242.28 | 348.89 |
|  | Overall Accuracy | 0.9520 | 0.9520 | 0.9432 | 0.9345 | 0.9389 |
| Stepwise | features | Var17,20,23,28 | Var8,20,23,28 | Var8,15,16,18,23,28 | Var17,20,21,23,26,28 | Var4,17,20,21,23,25 |
|  | C | 425.99 | 22.64 | 10.92 | 348.89 | 242.28 |
|  | Overall Accuracy | 0.9913 | 0.9869 | 0.9825 | 0.9825 | 0.9869 |

The experiment results are shown in Table 2. The experiments include three groups. In the first group, we used the whole features to build risk forecasting models as the threshold. And in the second group and the third group, we build risk forecasting models utilizing F-score and forward-back ward stepwise method to select features.

To evaluate the settlement risk forecasting models, we use five-folds cross validation overall accuracy. In each group, the overall accuracy decrease with the observation period becoming longer. However, the change of the overall accuracy is not obvious.

The features selected by F-score method are partly different from the one selected by forward-backward stepwise method. Take the forecasting model in advance 5 minutes for example, the features selected by F-score method are the range of the turnover, the variance of the price of the main contract, the range of the price of the main contract, the range of the net open rate and the variance of the turnover. While the different features selected by forward-backward stepwise method are the turnover in last t time and rate of the main contract in last t time.

## 6. Conclusion

Finally, it is useful to summarize what we have found and discuss in this paper. We successfully transfer the problem of settlement risk forecasting into a problem of classification. In order to build risk forecasting model, we utilize the standard SVMs based on two features selection method.

During the risk forecasting model building process, we attempt to find more powerful predictor no matter from economic meaning and from statistic significance. If we use overall accuracy as the criteria to evaluate risk forecasting model, we can conclude that the features selected by forward-backward stepwise have more prediction ability than the one selected by F-score method. However, the features selected by both feature selection methods in this paper are preliminary results. If we want to find the risk predictors in theory, it needs more expert experiences and further numerical experiments.

Further, through the numerical experiment, it demonstrates that the latest information has more prediction power.

## 7. Acknowledgement

This research has been partially supported by a grant from National Natural Science Foundation of China (#70621001,#10601064,#71071148,#70921061,#71110107026,#70933003). The authors thank for colleges Xi Zhao and Yibing Chen help with data pre processing.

Reference

1. Cont, R., *Statistical modeling of high-frequency financial data.* Signal Processing Magazine, IEEE, 2011. **28**(5): p. 16-25.
2. Engle, R.F. and M.E. Sokalska, *Forecasting intraday volatility in the US equity market. Multiplicative component GARCH.* Journal of Financial Econometrics, 2012. **10**(1): p. 54-83.
3. Brownlees, C.T., F. Cipollini, and G.M. Gallo, *Intra-daily volume modeling and prediction for algorithmic trading.* Journal of Financial Econometrics, 2011. **9**(3): p. 489.
4. Zhang, X. and Z. Chen. *Margin-Setting in Chinese Commodity Futures Markets: A VaR Approach*. 2008: IEEE.
5. Luo, J., et al., *The effect of differentiated margin on futures market investors' behavior and structure: An experimental research.* China Finance Review International, 2011. **1**(2): p. 133-151.
6. Liu, J., Y. Zhang, and S.Y. Li. *The Research of Stock Index Futures Margin Setting Model Based on Wavelet Analysis Method and Technology*. 2011: IEEE.
7. Wu, M. and S. Pang. *Stock index futures margin level settings by Hill estimation and empirical analysis*. 2010: IEEE.
8. Xie, H. and J. Li, *Intraday Volatility Analysis on S&P 500 Stock Index Future.* International Journal of Economics and Finance, 2010. **2**(2): p. P26.
9. Vapnik, V.N., *The nature of statistical learning theory*. 1995, Berlin: Springer.
10. Trafalis, T.B. and H. Ince. *Support vector machine for regression and applications to financial forecasting*. in *INNSENNS Int. Joint Conf. Neural Netw*. 2000. Italy: IEEE.
11. Tian, Y., Y. Shi, and X. Liu, *Recent advances on support vector machines research.* Technological and Economic Development of Economy, 2012. **18**(1): p. 5-33.
12. Noble, W.S., *Support vector machine applications in computational biology.* Kernel methods in computational biology, 2004: p. 71-92.
13. Borgwardt, K.M., *Kernel Methods in Bioinformatics.* Handbook of Statistical Bioinformatics, 2011: p. 317-334.
14. Ding, Z., *Application of Support Vector Machine Regression in Stock Price Forecasting.* Business, Economics, Financial Sciences, and Management, 2012: p. 359-365.
15. Lu, C.J., T.S. Lee, and C.C. Chiu, *Financial time series forecasting using independent component analysis and support vector regression.* Decision Support Systems, 2009. 47(2): p. 115-125.
16. Pearce, J. and S. Ferrier, *An evaluation of alternative algorithms for fitting species distribution models using logistic regression.* Ecological Modelling, 2000. 128(2): p. 127-147.
17. Menard, S.W., *Applied logistic regression analysis*. Vol. 106. 2002: Sage Publications, Inc.
18. Allison, P.D., *Logistic regression using the SAS system: theory and application*. 1999: SAS Publishing.

1. \* Corresponding author [↑](#footnote-ref-2)