Application of Unbalanced Data Classification Algorithm in Quantitative Financial Risk Management

JinPeng Zhu*, HanChen Wang

1Jiangsu University of Science and Technology, Zhenjiang, Jiangsu, China, 212003
2RMIT UNIVERSITY, 124 La Trobe St, Melbourne VIC 3000
*Corresponding author e-mail: MuShengdesu@gmail.com

Abstract. The global financial crisis that has erupted many times in recent years has not only harmed the financial systems of various countries, but even severely hit the economies of various countries. Therefore, the importance of financial risk early warning has become more prominent and the challenges facing it have become more severe. In view of this, it is very important to explore an extreme risk early warning model that suits the reality of the Chinese financial market, and to accurately predict and prevent extreme financial risks. This article first constructs the model's early-warning indicator system, and determines the state indicators by synthesizing the two state indicator definition methods based on the crisis period and EVT, in order to determine whether extreme financial risks occur in the financial market at a certain time. The prediction performance of the improved SVM under different unbalanced sample data sets is compared. It is highly feasible to use it for early warning of extreme risks in China's financial market. This paper presents a BADASYN algorithm based on boundary sample adaptive synthesis at the data level. The algorithm first finds a small number of samples in the class boundary region, then adaptively synthesizes some samples according to their distribution, and adds the newly synthesized samples to the training set. In the data set sampled by BADASYN, the support vector of the trained SVM model is mainly composed of newly synthesized samples, and finally the separation hyperplane is close to multiple types of samples. Experimental research shows that after testing SVMs constructed with four kernel functions, the prediction accuracy of each model is very high, reaching more than 93%, reflecting the stability of SVM prediction performance.

Keywords: Financial Risk, Intelligent Early Warning, Unbalanced Data Classification Algorithm, SVM Model

1. Introduction
In financial risk management, due to the impact of extreme fluctuations in the financial market, the extreme financial risk events that are triggered are often the focus of attention of financial and economic management departments and investors[1-2]. The reason is that although general risks appear frequently in the financial markets, they will not cause fatal threats. The extreme financial risks caused
by the extreme decline of the financial markets may cause the national economy to collapse and cause catastrophic consequences\cite{[3-4]}. It can be seen from this that exploring an extreme risk early warning model that suits the reality of the Chinese financial market and accurately predicting extreme financial risks is of great practical significance for investors to formulate investment strategies and prevent financial risks for financial and economic management departments\cite{[5-6]}.

For the problem of unbalanced data classification, there have been many conferences and workshops with the highest academic level in the past two decades. The number of papers on this issue has also increased year by year. Scholars at home and abroad have proposed many Ideas and learning algorithms\cite{[7-8]}.

Algorithm-level methods are based on the defects of traditional classification algorithms in solving unbalanced data distribution, and the algorithm is appropriately modified to adapt to the problem of unbalanced data classification, including cost-sensitive learning methods, integrated learning methods, and feature selection methods. The cost-sensitive learning method endows two types of samples with different misclassification costs, forcing the decision classifier to have a higher recognition rate for fewer types of samples\cite{[9-10]}.

This article builds an improved early warning model of extreme financial risks in China, and uses this to carry out early warning of extreme financial risks. It provides scientific reference for financial risk management departments to monitor extreme risks and investors to make investment decisions. With realistic value. This paper introduces the combination of non-equilibrium sample processing methods and traditional SVM to construct an improved SVM intelligent warning model for extreme financial risks in China, and uses this to carry out early warning of extreme financial risks, and strives to supervise extreme risks and investments for financial risk management departments. Investors make investment decisions to provide application tools with high operability. This paper proposes an integrated SVM learning method based on negative correlation learning and AdaBoost algorithm at the algorithm level. The simulation results on the UCI unbalanced data set show that: compared with the traditional negative correlation integrated learning algorithm and AdaBoostSVM algorithm, the proposed method has higher classification accuracy and better generalization ability.

2. Application of non-equilibrium data classification algorithm in quantitative financial risk management

2.1. Determination of Status Indicators

The alternative hypothesis is that the indicator can distinguish between extreme and non-extreme financial risk samples significantly. Then, compare the P value obtained from the independent sample T test with a certain significance level \( \alpha \). If the P value is less than \( \alpha \), the null hypothesis is rejected, indicating that at the \( \alpha \) significance level, the indicator can test extreme and non-extreme financial risk samples. Significantly distinguish, but not significantly distinguish; for indicators that do not obey the normal distribution, use the KS test method. The assumption conditions of the K-S test are unchanged, and the P value obtained by the K-S test can be judged for significance according to the above judgment method. It can be seen that the above-mentioned methods can further extract the internal risk characteristic indicators that have a significant impact on the explosion of extreme financial risks.

2.2. SVM-based Unbalanced Data Classification Algorithm

(1) Sampling-based method

Data-level methods are widely used in SVM-based unbalanced data classification algorithms. Before training the SVM model, the training samples are balanced by using various data preprocessing methods. These methods include: random up/down sampling, SMOTE sampling, and ADASYN algorithm. The SVM constructed on the basis of these sampling methods are: random sampling + SVM method, SMOTE sampling + SVM method, and various SVM methods based on the improved SMOTE algorithm.

1) Cost-sensitive support vector machine
The cost of misclassification for negative samples is $C^-$, where $C^+ > C^-$. That is to reduce the impact of data distribution imbalance on SVM performance by giving a higher misclassification cost to fewer samples.

$$\min \left( \frac{1}{2} |w|^2 + C^+ \sum_{i|y_i = +1} \xi_i + C^- \sum_{i|y_i = -1} \xi_i \right)$$  \hspace{1cm} (1)

2) zSVM

The zSVM proposed in this paper first uses the original unbalanced data set to train to obtain an SVM model; and then adjusts the decision boundary of the model to bias it to multiple types of samples. The adjusted decision function is shown in equation (2).

$$f(x) = s\text{ign}(z \ast \sum_{i=1}^{N^+} a_i^+ y_i K(x_i, x) + \sum_{i=1}^{N^-} a_i^- y_i K(x_i, x) + b)$$  \hspace{1cm} (2)

Among them, $a_i^+$ and $a_i^-$ are the coefficients of positive and negative support vectors, respectively, and $N^+$ and $N^-$ are the number of positive and negative samples, respectively. zSVM multiplies a $i^{th}+1$ by a particularly small positive number $z$ to increase the weight of the positive support vector, thereby moving the classification hyperplane to multiple types of samples.

(2) Evaluation method based on G, F and AUC

First define $|TS_{\text{min}}|$ and $|TS_{\text{maj}}|$ are the number of wrongly divided non-extreme risk samples into extreme risk samples and extreme risk samples into non-extreme risk samples, the same applies $|TS_{\text{min}}|$ and $|TS_{\text{maj}}|$ are the number of extreme risk samples and non-extreme risk samples that are correctly classified.

The $F_{\text{measure}}$ of extreme risk samples is a commonly used evaluation index in the prediction of non-equilibrium data. It focuses more on the classification performance of extreme risk samples. Only when the extreme risk sample precision rate SP and extreme risk correct rate SE are both large, the extreme risk sample $F_{\text{measure}}$ will increase. Therefore, it can accurately reflect the prediction performance of the model for extreme risk samples.

In addition, the AUC value is another widely used performance evaluation index of the unbalanced data set model, which is used to evaluate the comprehensive prediction performance of two types of samples (Marroco et al., 2008). The AUC value represents the area under the ROC curve, and the early warning model corresponding to the large AUC value is more superior to the comprehensive prediction performance of the two types of samples.

3. Early warning indicator system construction plan

3.1. Selection of Samples

For the study of extreme risks in China's financial market, the selection of the research object needs to be representative. This article selects such a long range from January 2017 to February 2018, a total of more than 2,000 trading days of data as the research object because the Chinese stock market has experienced the entire process from a sharp rise to a sharp fall during this period, and this range also It fully covers the global financial crisis that has occurred in recent years. The research samples selected in this way will make the prediction results of the extreme risk crisis early warning model more convincing.

3.2. SVM Risk Prediction Performance Evaluation Method

After the risk warning model is constructed, in order to test whether the model is constructed successfully, there must be a scientific evaluation method. Only by passing the test of scientific evaluation method can we ensure the practicability of the model. At present, most of the papers use the prediction accuracy as the standard when evaluating SVM prediction performance. The calculation method of the prediction accuracy Acc is as follows:

$$\text{Acc} = \frac{|T|}{|T| + |F|}$$  \hspace{1cm} (3)

Among them, $|T|$ is the number of samples with correct prediction, and $|F|$ is the number of samples with incorrect prediction. It can be seen that the model's prediction accuracy Acc can reflect the
model's overall prediction accuracy for all samples, and has a certain effect on the model's prediction performance evaluation.

3.3. Convergence Experiment of the Algorithm
Since the number of iterations of the algorithm in this paper is mainly determined by the update step $\sigma_{step}$ of the Gaussian kernel width $\sigma$ of RBF SVM, this paper adjusts the number of iterations by adjusting $\sigma$ step. The relationship between the training error of Ionosphere and Soybean data sets and the number of iterations is given. It can be seen that with the increase of the number of iterations, the training accuracy gradually improves and finally stabilizes. At the same time, it can be seen from the experimental results that when the number of iterations is equal to 20, the accuracy has tended to be stable, so the step $\sigma_{step}$ is set to 2 in the experiment.

4. Experimental analysis of non-equilibrium data classification algorithm in quantitative financial risk management

4.1. Evaluation Method Effect Test
In order to test the evaluation method using G, F and AUC as the standard used in this paper, the SVM model constructed above will be tested, with CSI300 as the research object, and the two types of samples identified in this paper will be used as test data to find the SVM The prediction results under different kernel functions are shown in Table 1. The entire test is based on Matlab2013b programming software.

| Kernel function | Acc(%) | $G_{mean}$ | $F_{measure}$ | AUC   |
|------------------|--------|------------|---------------|-------|
| Linear           | 93.744 | 0          | NaN           | 0.7396|
| Polynomial       | 93.563 | 0          | NaN           | 0.7351|
| RBF              | 93.582 | 0          | NaN           | 0.7322|
| Sigmoid          | 93.275 | 0          | NaN           | 0.7382|

By observing Table 1, we can find that the prediction accuracy of SVM is relatively high. After testing the SVM constructed with four kernel functions, the prediction accuracy of each model is very high, reaching more than 93%. If the prediction accuracy Acc is used as the basis for judging the performance of the model, no matter which kernel function is used to construct the SVM, it has a good prediction effect, and there is almost no difference between the prediction results of the SVM using different kernel functions. It also reflects the stability of SVM's predictive performance. However, by observing the values of G, F and AUC in Table 1, the opposite conclusion can be obtained. Under the four kernel functions, the G value of the SVM model is 0, and the F value is NaN, which is caused by $|TS_{min}|$, the number of extreme financial risks that are correctly predicted is 0, and the AUC value is just over 0.6 However, it shows that the SVM model does not have a good prediction effect on a few samples, that is, extreme financial risks under non-equilibrium data.

Through the above comparison, we can see that using two different performance evaluation methods, the performance evaluation of the model is completely opposite, and the G, F and AUC values are used as the performance evaluation method of the extreme financial risk early warning model constructed in this paper. Will be able to make a more scientific and accurate evaluation of the model's predictive performance.

4.2. Comparison of Generalization Error
This paper uses the method of 10-fold cross-validation to estimate the generalization error and take the average value as the training error of the final classifier. Both AdaBoost-SVM and Bagging have 50 iterations, GASEN has 100 iterations, and NCCD has 20 iterations. The minimum classification error on each data set is shown in bold. The experimental results are shown in Figure 1 below.
It can be seen that NCAB-SVM obtains the smallest classification error in most data sets compared to other algorithms, indicating that the algorithm in this paper not only improves the performance of the AdaBoostSVM algorithm, but also introduces SVM as an integrated system compared to other negative correlation learning algorithms. The base classifier can improve the performance of the integrated system. At the same time, it can be seen that the bagging method based on sample perturbation has not significantly improved the classification accuracy compared to the pure SVM. Therefore, it is necessary to tailor the base classifier with redundant or useless information.

5. Conclusion
In order to improve the classification accuracy of SVM, this paper proposes an integrated learning method of SVM based on negative correlation learning and AdaBoost algorithm. The negative correlation learning theory is integrated into the training process of AdaBoostSVM, and the correlation between the base classifiers is calculated using the negative correlation learning theory, and the weights of the base classifier are adaptively adjusted according to the correlation value, and then the weighted decision classifier is obtained. This paper analyzes the shortcomings and deficiencies of SVM when classifying unbalanced data: Because SVM is based on the method of soft-space maximization, the classification hyperplane will be inclined to a few classes; the imbalance of support vector ratio will also lead to test samples Is surrounded by more negative support vectors. This paper believes that the Borderline-SMOTE-EasyEnsemble-SVM model can accurately predict the extreme risks of my country's financial market and has high practical value. For the financial and economic management departments, the model can be used to predict whether future extreme financial risks will occur more accurately, so as to formulate corresponding economic policies to resist the impact of financial risks, maintain financial order stability, and avoid the violent operation of the financial market. The volatility of the economy will eventually play a positive role in creating a good and stable environment for the rapid development of the macro economy; for financial participants, the model is used to predict the extreme risks of financial products, and it is more effective based on the prediction results Carry out risk management by adjusting financial asset investment strategies in a timely manner to avoid possible extreme financial risks, and then reduce losses and increase returns as much as possible during the investment and financial management process.

References
[1] Dalla Valle L , De Giuli M E , Tarantola C , et al. Default Probability Estimation via Pair Copula Constructions[J]. European Journal of Operational Research, 2016, 249( 1):298-311.
[2] Barot P A , Jethva H B . Unbalanced Data Classification using Feature Selection through
BitApriori Algorithm.[J]. INTERNATIONAL JOURNAL OF COMPUTER ENCES AND ENGINEERING, 2018, 6(10):701-704.

[3] Kang Q . Financial risk assessment model based on big data[J]. International Journal of Modeling Simulation & Scientific Computing, 2019, 10(04):106-113.

[4] Bartram S M , Brown G W , Waller W . How Important Is Financial Risk?[J]. Journal of Financial and Quantitative Analysis, 2015, 50(4):801-824.

[5] Antipina, O. V, Prokopyeva, A. C. CLASSIFICATION OF FINANCIAL RISKS AND MANAGEMENT TECHNIQUES IN THE ORGANIZATION OF PRODUCTION PROCESSES[J]. The World of entific Discoveries, 2015, 25(5.8):2982.

[6] Glau K , Scherer M , Zagst R . [Springer Proceedings in Mathematics & Statistics] Innovations in Quantitative Risk Management Volume 99 || A Variational Approach for Mean-Variance-Optimal Deterministic Consumption and Investment[J]. 2015, 10.1007/978-3-319-09114-3(Chapter 13):225-238.

[7] Sgouras K I , Dimitrelos D I , Bakirtzis A G , et al. Quantitative Risk Management by Demand Response in Distribution Networks[J]. IEEE Transactions on Power Systems, 2018, 33(2):1496-1506.

[8] Oppliger R . Quantitative Risk Analysis in Information Security Management: A Modern Fairy Tale[J]. IEEE Security & Privacy, 2015, 13(6):18-21.

[9] Hugo F D , Pretorius L , Benade S J . SOME ASPECTS OF THE USE AND USEFULNESS OF QUANTITATIVE RISK ANALYSIS TOOLS IN PROJECT MANAGEMENT[J]. South African Journal of Industrial Engineering, 2018, 29(4):116-128.

[10] Budetta P , De Luca C , Nappi M . Quantitative rockfall risk assessment for an important road by means of the rockfall risk management (RO.MA.) method[J]. Bulletin of Engineering Geology and the Environment, 2016, 75(4):1377-1397.