Forecasting of Credit Card Default Based on Incremental Random Forest

Xiang Gao¹,², Junhao Wen²,*, and Cheng Zhang¹

¹College of Computer Science, Chongqing University, Chongqing, China
²College of Big Data & Software Engineering, Chongqing University, Chongqing, China
³Chongqing Branch of China Minsheng Bank, Chongqing, China

*Email: jhwen@cqu.edu.cn

Abstract. In this paper, the incremental random forest algorithm is proposed for the classification and prediction problem of dynamically increasing data. Traditional batch machine learning algorithms perform modeling at one time and cannot allow newly generated samples to participate in learning, which leads to too much model deviation. This paper combines incremental learning with random forest and proposes incremental random forest. Applying this algorithm to the problem of predicting credit card customer default behavior can help banks control risks and reduce losses. It is important to conduct card issuance audits on card issuers and early warning of risks to cardholders. The algorithm performed better in the experiment of predicting the default behavior of credit card customers based on a batch of credit card holder data of a bank in Taiwan. Compared with random forest, decision tree, logistic regression, naive bayes, BP neural network, and support vector machine, it has relatively better performance in our experiment.

1. Introduction

Compared with ordinary bank savings cards, credit cards are convenient to use even if you don’t have cash. You can use your credit line for overdraft consumption without additional charge, as long as you return the overdraft amount on time before the repayment date. Therefore, more and more people use credit cards. For banks, vigorously developing credit card business can directly obtain profits from interest income, annual fee income, merchant rebate income, cash withdrawal fees and punitive fees, and some value-added service income. Therefore, major banks want do everything possible to maximize the development of customers in order to seize more market shares. Every bill of credit card has a minimum repayment amount. Once the repayment amount of your bill is lower than this amount, it means that you are overdue, that is, you are in default. There are some customers who may breach the contract for various reasons and damage the interests of the bank. They may be rascals who want to take advantage of the bank’s loopholes. They may have a big money gap due to sudden changes. They may spend more on high-end consumption without restraint than they earn. They may fail in fundraising and entrepreneurship and even several credit cards could not repay, and some people even used other people’s identity cards with bad motives to make purchases at the beginning. So banks urgently need to control defaults in order to prevent risks and reduce losses. This article predicts the default behavior of credit card applicants or cardholders, and provides banks with effective means of credit card approval and risk control.

Many scholars have proved that machine learning has great practical value in solving
various risk problems, including credit risk prediction. Luo et al. [1] proposed a clustering-based classification algorithm (CLC) and applied it to two real-world credit card datasets provided by the own machine learning database of the university of California. Practice has proved that CLC is more efficient than support vector machine. Wang et al. [2] found that the combined classifier and its corresponding integrated learning algorithm have obvious advantages in credit scoring. TWALA et al. [3] also used the combined classifier in credit risk prediction. They found that the prediction results were significantly improved compared with the classification performance of individual classifiers by voting on sub-classifiers of combined classifiers. Yap et al. [4] studied credit scoring models based on historical payment data and demographic characteristics using statistical techniques. This model can not only predict a person's risk label, but also find out the important characteristics that determine a person's credit risk state. Li et al. [5] introduced the idea of transfer learning. Based on the similarity between traditional business and new business, they migrate the data of traditional banking to new business data. Through experiments, they prove the commercial value of transfer learning concept in credit evaluation in the field of financial risk, and provide decision basis for practitioners and managers.

All the above studies are one-time batch learning techniques. The technique assumes that all training samples can be obtained at one time before training, and after learning these samples, the learning process terminates and no longer learns new knowledge. In the real world, however, data sets for modeling tend to accumulate and increase. Once the model is built, ignoring the newly arrived sample will not reflect the real situation, and the prediction results will deviate.

In order to bring the new data into the modeling process, this paper introduces the idea of incremental learning and proposes an incremental random forest [6] algorithm. The algorithm is used in credit card customer default behavior prediction to show its value.

2. Incremental Random Forest Algorithm
Incremental learning is a machine learning method [7]. It continuously applies the input data to expand the knowledge of existing models to further train the models. Through incremental learning, a learning system can constantly learn new knowledge from new samples and preserve most of the knowledge that has been learned before. Incremental learning is very similar to the human learning model. In the process of human growth, learning and receiving new things every day is carried out step by step. People will not forget what he has learned usually.

The importance of incremental learning is mainly reflected in two aspects:
1) For the actual perceptual data, the amount of data is often gradually increased. Therefore, learning methods should be able to make some changes to the trained system with input of new data in order to learn the knowledge contained in the new data.
2) The time cost of modifying a trained system is usually lower than the cost of retraining a system.

Incremental learning ideas are described below: Whenever you add data, you don't need to rebuild all the knowledge base. Instead, only the changes caused by the new data are updated on the original knowledge base. The incremental learning method is more in line with the principle of human thinking. There are many incremental learning frameworks, and the core content of each framework is the similarity evaluation method of processing new data and stored knowledge, because this method affects the growth of knowledge. Therefore, the judgment mechanism of new knowledge is the core component of incremental learning.

Compared with the traditional random forests that support batch learning, this paper proposes a random forest classification algorithm that supports incremental learning (not online learning). The idea is:
1) Based on the existing samples, the traditional random forest is used to establish the basic model. Unlike traditional random forest, the training samples need to be saved, that is, the split rules of intermediate nodes, the samples of leaf nodes, and the class of leaf nodes are saved.
2) The full amount of samples arriving over time will directly participate in incremental modeling. In the incremental learning scenario, subsequent samples arrive one after another over time, and the traditional self-service sampling strategy is no longer applicable. All arriving samples are arranged to be received by the base classifier. This method can reduce deviation.

3) Build a classification decision tree that supports incremental learning. First, the sample is stored at the leaf node. The node is then split when the impurity of the leaf node reaches the threshold. The nonpurity rule can be either the Gini coefficient or entropy. We choose the Gini coefficient to calculate the purity. The split threshold can be obtained by selecting the highest accuracy value of the split tree.

4) Discard the sample with the smallest contribution. According to the Least Recent Use principle, the sample set contribution on the leaf node with the earliest recent access time is the smallest. In order to adapt to the classification task in big data environment, it is necessary to formulate sample discard strategy to save the space occupied by the model.

In training, for newly arrived samples, they are first stored on the leaf nodes of the corresponding classification decision tree according to the constructed random forest and marked on the leaf nodes. The marked leaf nodes are then detected with certain rules to determine whether or not to split, and recursively split the leaves. Finally, determine whether the total amount of sample storage needs to trigger the discard strategy.

The pseudo code of the incremental random forest algorithm is as follows:

---

**Incremental Random Forest Algorithm**

**Input:**
- A newly arrived training sample \( s \);
- A trained completed random forest with a tree (first iteration) or incremental random forest \( T \)

**Output:**
- An incremental random forest with \( K \) trees

1: For decision tree \( T_i \) in \( T \) do

1.1: Place \( s \) from the root node of the current incremental classification tree \( T_i \) to the corresponding leaf node \( L_i \) according to the split rule

1.2: Store \( s \) in the leaf node \( L_i \)

1.3: Record the total number of samples of the tree \( totals_i + 1 \)

1.4: if Unpurity of leaf nodes \( L_i \) \( I(L_i) > a \) then

// Here gini(Li) is used for splitting judgment, assumption \( a = 0.5 \)

1.5: Split the leaf nodes \( L_i \) and get the left and right nodes

1.6: Repeat 1.4 to 1.10 for the left node

1.7: Repeat 1.4 to 1.10 for the right node

1.8: else

1.9: Marking time stamps for leaf nodes \( L_i \)

---
1.10: endif
1.11:  If \( \text{totals}_i > \text{MaxSamples} \) then
1.12:   Discard the sample set \( D_i \) stored by the earliest leaf nodes according to the Least Recent Use principle;
1.13:   \( \text{totals}_i = \text{totals}_i - \text{sizeof} \ (D_i) \);
1.14:  endif
End For

2: Output incremental random forest

\( T' \)

3. Experiments
The experimental data of this paper are derived from the basic information of a group of credit card holders in a bank in Taiwan, the credit line, the billing information and repayment information for a total of six months from April to September 2005, and the default Label. To simulate the real world, we assume that samples with ID number less than or equal to 5000 arrive at one time, and the remaining 25000 samples arrive in order of ID number.

Because random forests have the advantage of processing high-dimensional data, we no longer reduce the dimension of features. However, we can obtain the importance ranking of features in training and excavate key elements for risk control. There are 23 input variables and 1 output variable, that is, category attribute.

In the experiment, we use Python to sort the feature importance and weka to achieve performance comparison between incremental random forest algorithm and other classical algorithms.

Using random forest to sort the importance of the feature, we can see that the most significant variable in the prediction of credit card customer default behavior involved in this data determines whether the customer defaults or not: “repayment status of September 2005(last month)”, “age”, “bill amount of September 2005”, “credit lines”.

| Algorithm   | ACC (%) | Precision | Recall | AUC  | Time(s) |
|-------------|---------|-----------|--------|------|---------|
| RF          | 81.56   | 0.797     | 0.816  | 0.767| 37.8    |
| C4.5        | 80.32   | 0.782     | 0.782  | 0.658| 8.56    |
| Logistic    | 81.03   | 0.795     | 0.810  | 0.723| 3.26    |
| NaiveBayes  | 69.35   | 0.770     | 0.694  | 0.733| 0.22    |
| BP          | 81.68   | 0.799     | 0.817  | 0.743| 216.65  |
| SVM         | 80.93   | 0.793     | 0.809  | 0.604| 591.27  |
| IRF         | 81.81   | 0.801     | 0.818  | 0.790| 25.13   |

Table 1. Experimental Results of Different Algorithms.
Our experiment shows that incremental random forest (IRF) has better performance than random forest (RF), decision tree (C4.5), logistic regression (Logistic), naive bayes [8] (NaiveBayes), BP neural network (BP), and support vector machine [9] (SVM) algorithm. It can be found from Table 1 that the IRF has better performance than the other six algorithms in the four evaluation indexes of accuracy, precision, recall and AUC (Area Under Curve). The results show that the IRF has better performance than the other six algorithms. The modeling time (Time) is shorter than RF, BP and SVM. Figure 1 is a curve comparison of 5 of the 7 algorithms with larger AUC values. The curve of the IRF is closer to the Y axis, with the largest area under it, followed by RF.

4. Conclusion
In this paper, incremental random forest prediction method is proposed by combining incremental learning with random forest. The experiment of credit card data of a bank in Taiwan shows that incremental random forest has better performance than other traditional methods in predicting credit card customer default behavior. This method is not only suitable for risk prediction, but also has some advantages for all dynamically increased data sets.

Acknowledgments
The work described in this paper was supported by the National Natural Science Foundation of China (Grant no. 61672117).

References
[1] Luo S -T, Cheng B –W 2009 Prediction model building with clustering-launched classification and support vector machines in credit scoring Expert Systems with Applications 36(4) 7562-6
[2] Wang G, Hao J, Ma J, et al. 2011 A comparative assessment of ensemble learning for credit scoring Expert Systems with Applications 38(1) 223-230
[3] Twala, Bhekisipho 2009 Combining classifiers for credit risk prediction Journal of Systems Science and Systems Engineering 18(3) 292-311
[4] Wah Y B, Huat O S, Husain 2011 Using data mining to improve assessment of credit worthiness via credit scoring models Expert Systems with Applications 38(10) 13274-83
[5] Ding L W, Chen S, Wang W, et al 2001 Transfer learning-based default prediction model for consumer credit in China The Journal of Supercomputing 75(2) 862-84
[6] Breiman L 2001 Random forests *Machine Learning* 45(1) 5-32
[7] Gu B, Sheng V S, Tay K Y, et al 2014 Incremental support vector learning for ordinal regression *IEEE Transactions on Neural Networks and Learning Systems* 26(7) 1403-16
[8] Zhang H, Liu G, Chow T W S, et al 2011 Textual and visual content-based anti-phishing: A bayesian approach *IEEE Transactions on Neural Networks* 22(10) 1532-46
[9] Cortes C and Vapnik V 1995 Support-vector networks *Machine Learning* 20(3) 273–297