Research on Personal Credit Risk Assessment Model Based on Instance-Based Transfer Learning

Maoguang Wang, Hang Yang

Research on Personal Credit Risk Assessment Model Based on Instance-based Transfer Learning

Maoguang Wang[1] , Hang Yang[1]

[1] Central University of Finance and Economics, School of Information，China
{Mgwangtiger,Yanghangv}@163.com

Abstract. Personal credit risk assessment is an important part of the development of financial enterprises. Big data credit investigation is an inevitable trend of personal credit risk assessment, but some data are missing and the amount of data is small, so it is difficult to train. At the same time, for different financial platforms, we need to use different models to train according to the characteristics of the current samples, which is time-consuming. In view of these two problems, this paper uses the idea of transfer learning to build a transferable personal credit risk model based on Instance-based Transfer Learning(Instance-based TL). The model balances the weight of the samples in the source domain, and migrates the existing large dataset samples to the target domain of small samples, and finds out the commonness between them. At the same time, we have done a lot of experiments on the selection of base learners, including traditional machine learning algorithms and ensemble learning algorithms, such as decision tree, logistic regression, xgboost and so on. The datasets are from P2P platform and bank, the results show that The AUC value of Instance-based TL is 24% higher than that of the traditional machine learning model, which fully proves that the model in this paper has good application value. The model's evaluation uses AUC、prediction、recall、F1. These criteria prove that this model has good application value from many aspects. At present, we are trying to apply this model to more fields to improve the robustness and applicability of the model; on the other hand, we are trying to do more in-depth research on domain adaptation to enrich the model.

Keywords: Personal Credit Risk, Big Data Credit Investigation, Instance-based Transfer Learning.

1 Introduction

Personal credit risk is a part that both government and enterprises attach great importance to. A good personal credit risk assessment will not only help government to improve the credit system but also make some enterprises avoid risk effectively. The development of personal credit risk assessment model is from traditional credit assessment model to data mining credit risk assessment model. It has gone through the process from traditional credit assessment model to big data credit assessment model. Traditional credit assessment model often uses

discriminant analysis, liner regression, logistic regression, while data mining credit risk assessment model often use decision tree, neural network, support vector machine and other methods to evaluate credit[1].

At present, the existing data mining credit risk assessment models have relatively high accuracy, but only limited to the case of sufficient data and less missing values. When the data volume is small or the data is seriously missing, the prediction effect of the model is often poor. Based on this, we introduces Instance-based Transfer Learning, which migrates the existing large data set samples to the target field of small samples, finding out the commonness between them, and realizing the training of the target domain dataset.

In the other parts, the second section introduces the related works. The third section constructs the personal credit risk assessment model based on the idea of Instance-based transfer. The fourth section introduces the specific experimental process and the comparative analysis of the results. The fifth section is the summary of the full paper.

2 Related Works

The concept of transfer learning was first proposed by a psychologist. Its essence is knowledge transfer and reuse. Actually, it is to extract useful knowledge from one or more source domain tasks and apply it to new target task, so as to realize “renovation and utilization” of old data and achieve high reliability and accuracy. The emergence of transfer learning solves the contradiction between “big data and less tagging” and “big data and weak computing” in machine learning.

In terms of the classification of transfer learning, Pan, S. J. and Yang, Q.[2] summarized the concept of transfer learning and divided transfer learning in 2010 according to learning methods ,which can be divided into Instance-based Transfer Learning, Feature based Transfer Learning, Model based Transfer Learning and Relation based Transfer Learning. According to the characteristic attributes, transfer learning can also be divided into Homogeneous Transfer Learning and Heterogeneous Transfer Learning[3]. According to the offline and online learning system, transfer learning can also be divided into Offline Transfer Learning and Online Transfer Learning. Among these classification, Instance-based Transfer Learning is the most commonly used model. The Instance-based Transfer Learning generate rules based on certain weights and reuse the samples. Dia et al.[4] proposed the classic TrAdaboost method, which is to apply the AdaBoost idea into transfer learning. It is used to increase the weight beneficial to the target classification task and reduce the weight harmful to the classification task,so as to find out the commonness between the target domain and the source domain and realize the migration.

At present, transfer learning has a large number of application, but mainly concentrated in Text Classification, Text Aggregation, Emotion Classification, Collaborative Filtering, Artificial Intelligence Planning, Image processing, Time Series, medical and health fields.[5-7] Dia et al.applied Feature based Transfer

Learning to the field of text classification and achieved good results[8]. Zhu et al. [9] proposed a Heterogeneous Transfer Learning method in the field of image classification. Pan et al.[10] applied transfer learning algorithms to Collaborative Filtering. Some scholars also use transfer learning framework to solve problems in the financial fields. Zhu et al.introduced TrBagg which can integrate internal and external information of the system, in order to solve the category imbalance caused by the scarcity of a few samples in customer credit risk[11].Zheng, Lutao et al. improved TrAdaBoost algorithm to study the relationship between user behavior and credit card fraud[12]. Wang Xu et al. applied the concept of migration learning to quantitative stock selection[13]. But generally speaking, transfer learning is seldom used in the financial field, especially in the field of personal credit risk.

3 The Construction of Personal Credit Risk Assessment Model

3.1 The Build of Instance-based Transfer Learning

Traditional machine learning assumes that training samples are sufficient and that training and test sets of the data are distributed independently. However, in most areas, especially in the field of financial investigation, these two situations are difficult to meet, data sets in some domains have not only small data volume but also a large number of missing, which leads to the traditional machine learning method can not train very good results. If other data sets are introduced to assist training, it will be unable to train because of the different distribution of the two data sets. In order to solve this problem, we introduces Transfer Learning. In transfer learning, we call the existing knowledge or source domain, and the new knowledge to be learned as the target domain. And Instance-based Transfer Learning, to make maximum use of the effective information in the source domain data to solve the problem of poor training results caused by the small sample size of the target domain data set.

In order to ensure the maturity of the transfer learning framework， we innovatively introduce the classic algorithm of Instance-based Transfer Learning, the tradabost algorithm, to apply to the data in the field of financial credit reference [4].The TrAdaBoost algorithm comes from the Ensemble Learning-AdaBoost algorithm, which is essentially similar to the AdaBoost algorithm. First of all, it gives weight to all samples, and if a sample in the source data set is misclassified during the calculation process, we think that the contribution of this sample to the destination domain data is small, thus reducing the proportion of the sample in the classifier. Conversely, if the sample in the destination domain is misclassified, we think it is difficult to classify this sample, so we can increase the weight of the sample. The sample migration model built in this paper is based on the tradaboost framework, which is divided into two parts: one is the construction of the tradaboost framework [4,14,15], the other is the selection of the relevant base Learners [16]. The following figure shows specific process.

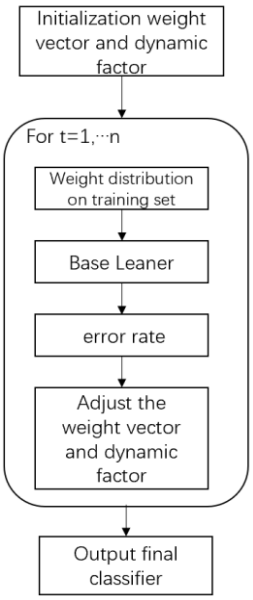

Fig. 1. Instance-based TL

Where we mark the source domain data as T_a, the destination domain data is marked T_b. Take 50% of all the source domain data and the target domain data as the training set T, take 50% of the target domain data as the test set, recorded as S, from which it is not difficult to find that T_b and S are same distribution.

Step 1. Normalized training set($T_a \cup T_b$) And each data weight in the test set (S) to make it a distribution.

Step 2. For t=1,…,N

(1)Set and call the Base Learner.

(2)Calculate the error rate, and calculate the error rate on the training set S.

(3)Calculates the rate of weight adjustment.

(4)Update the weight. If the target domain sample is classified incorrectly, increase the sample weight; if the source domain sample is classified incorrectly, reduce the sample weight.

Step 3. Output final classifier

3.2 Base Learner Selection

In general, for personal credit risk assessment, the commonly used algorithms include logistic regression, decision tree and other machine learning algorithms, as well as xgboost and other Ensemble Learning algorithms. When the dataset is sufficient, the application of machine learning algorithm on the dataset can achieve good results. Therefore, we can learn from these mature algorithms in the selection of Base Learner, and migrate the algorithm from the source domain to the target

domain, so as to achieve better results in the target domain.

Learners are generally divided into weak learners and strong learners. At present, most researches choose weak learners, and then through many iterations to achieve better results. However, we find that in the field of credit risk, some scholars have applied xgboost algorithm and achieved good results[16] . Therefore, according to the characteristics of data in the field of credit risk, this paper selects the strong learner-xgboost algorithm as the Base Learner, which is also convenient for model parameter adjustment and optimization.

XGBoost (extreme gradient boosting) is a kind of Ensemble Learning algorithm, which can be used in classification and regression problems, based on decision tree. The core is to generate a weak classifier through multiple iterations, and each classifier is trained on the basis of the residual of the previous round. In terms of prediction value, XGBoost's prediction value is different from other machine learning algorithms. It sums the results of trees as the final prediction value.

$$\hat{y}_i = \emptyset(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{1}$$

Suppose that a given sample set has n samples and m features, which is defined as

$$\mathrm{D} = \{ （ x_i^-, y_i ） \} （ |\mathrm{D}| = \mathrm{n}, x_i \in R^m, y_i^- \in \mathrm{R} ） \tag{2}$$

For x_i ,y_i ,The space of CART tree is F. As follows:

$$\mathrm{F} = \{\mathrm{f(x)} = w_{q(x)}\}(\mathrm{q}: R^m \to \mathrm{T}, \mathrm{w} \in R^T) \tag{3}$$

Where q is the model of the tree, $w_{q(x)}$ is the set of scores of all leaf nodes of tree q; T is the number of leaf nodes of tree q. The goal of XGBoost is to learn such k-tree model f (x). Therefore, the objective function of XGBoost can be expressed as[13]:

$$obj^t = \sum_{i=1}^{n} l(y_j, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_t) \quad where\ \Omega(f) = \Upsilon T + \frac{1}{2}\lambda \|w\|^2 \tag{4}$$

4 Compare Experiments and Results Analysis

4.1 The Source of the Dataset

The source domain dataset and target domain dataset are from the Prosper online P2P lending website and a bank's April-September 2005 , respectively. The data sets of both source domain and target domain data have data missing and high correlation among features. There are only 9000 pieces of data in the destination domain, and the source domain dataset contains more redundant fields. Therefore, it is necessary to fill in the missing values and select features by information divergence.

4.2 Missing Values Processing

There are several common missing value handling methods:
(1) Filling fixed values according to data characteristics;
(2) Fill the median/median/majority;
(3) Fill in the KNN data;
(4) Fill the predicted value of the model;

4.3 Feature Selection

The characteristics of the data will have a positive or negative impact on the experimental results. In particular, the amount of features in the source domains of this paper is huge, including many redundant features and highly relevant features. Firstly, delete redundant features according to the meaning of the features. The following table is the feature dictionary after the features are deleted.

Table 1. Deleted characteristic values.

Common ground	Features	Meaning
redundant features	ListingKey	Unique key for each listing, same value as the 'key' used in the listing object in the API.
	ListingNumber	The number that uniquely identifies the listing to the public as displayed on the website.
	LoanNumber	Unique numeric value associated with the loan.
	LenderYield	The Lender yield on the loan. Lender yield is equal to the interest rate on the loan less the servicing fee.
	LoanKey	Unique key for each loan. This is the same key that is used in the API.
Characteristics related only to investors	LP_InterestandFees	Cumulative collection fees paid by the investors who have invested in the loan.
	LP_CollectionFees	Cumulative collection fees paid by the investors who have invested in the loan.
	LP_GrossPrincipalLoss	The gross charged off amount of the loan.
	LP_NetPrincipalLoss	The principal that remains uncollected after any recoveries.
	PercentFunded	Percent the listing was funded.
	InvestmentFromFriendsCount	Number of friends that made an investment in the loan.
	InvestmentFromFriendsAmount	Dollar amount of investments that were made by friends.

We chooses the method of Information divergence to select other features. Information divergence is often used to measure the contribution of a feature to the whole, which also can select features. The basis of Information divergence is entropy, a measure of the uncertainty of random variables. Entropy can be subdivided into information entropy and conditional entropy. The computational formula is shown in Table 2[17].

Table 2. Calculation formula of entropy.

Information Entropy	$H(S) = -\sum_{i=1}^{C} p_i log_2(p_i)$
conditional entropy	$H(C\|T) = P(t)H(C\|t) + P(\bar{t})H(C\|\bar{t})$

The calculation of Information divergence is based on information entropy and conditional entropy. The computational formula is as follows.

$$IG(T) = H(C) - H(C|T) \tag{5}$$

Using the python program, the entropy of the overall dataset and Information divergence of each feature can be obtained. At the same time, the greater the value of Information divergence, the greater the contribution of the feature to the overall dataset. Since there are many useless features in the source domain data in this paper, Information divergence of each feature is calculated and shown in Fig. 2. In order to keep consistent with the target domain, this paper selects the first 23 features with greater Information divergence to simplify the subsequent calculation process.

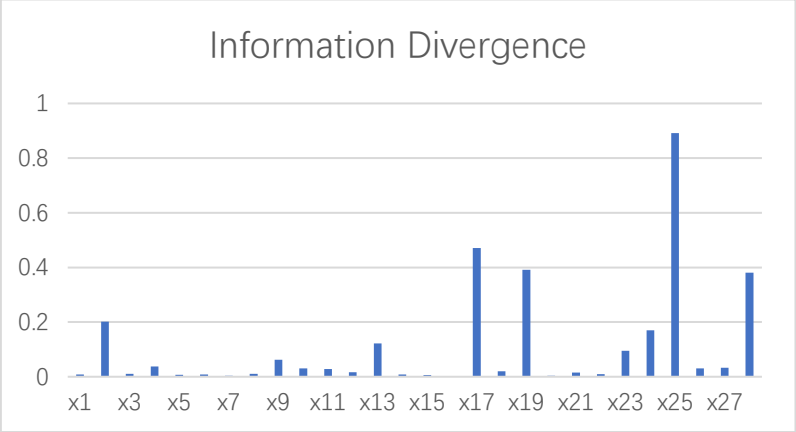

Fig. 2. Source domain characteristic information divergence

4.4 Experimental Results and Comparative Analysis Results

Firstly, apply the XGBoost algorithm to training T_a and T_b. The training results are as follows

Table 3. Training results.

Dataset	T_a	T_b
AUC	0.97	0.56

It is observed that training T_b alone cannot get a better performance. However, using the XGBoost algorithm to train T_a can get a higher AUC value, which

proves that it is feasible to use the XGBoost experimental method as a Base Learner.

In the aspect of base learner selection, we have done a lot of experiments, including traditional machine learning algorithm and ensemble learning algorithm. In this paper, we choose xgboost as the base learner to construct the tradapoost (xgboost). The following table shows the experimental results. Fig. 3. shows the AUC value of the tradapoost (xgboost).

Table 4. The results of tradapoost (xgboost)

	AUC	prediction	recall	F1
TrAdaBoost(XGBoost)	0.80	0.79	0.65	0.71

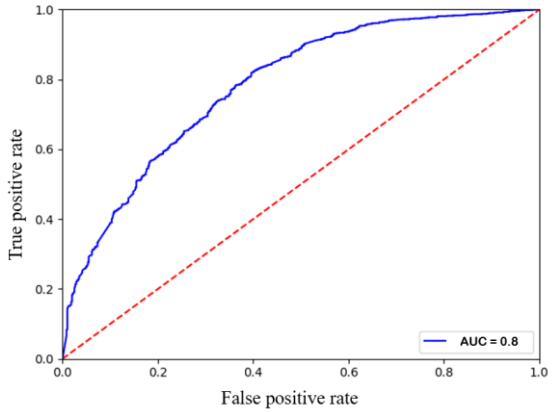

Fig. 3. The AUC of the tradapoost (xgboost)

It can be seen that the accuracy of T_b after transfer is significantly higher than that of training using only XGBoost algorithm.

The experiment is compared from two aspects :(1) Choose different Base Learners, and compare it from the transfer learning dimension.(2) Compare transfer learning with machine learning algorithms.

In the dimension of transfer learning, this paper adds the decision tree as the Base Learner of TrAdaBoost to predict data. Denote the algorithm using decision tree as the base learner as TrAdaBoost（DT）.At the same time, Denote the algorithm using XGBoost as the base learner as TrAdaBoost（XGBoost）.Now, we input into the Base Learner using decision tree and XGBoost as TrAdaBoost construction separately to predict the target data. The AUC value is selected as the criterion of result evaluation. The models' evaluation uses AUC、prediction、recall、F1. Table 5 and Figure 4 show the results.

Table 5. Comparison results-1.

TL VS TL	TrAdaBoost（DT）	TrAdaBoost（XGBoost）

AUC	0.62	0.80
prediction	0.64	0.79
recall	0.61	0.65
F1	0.63	0.71

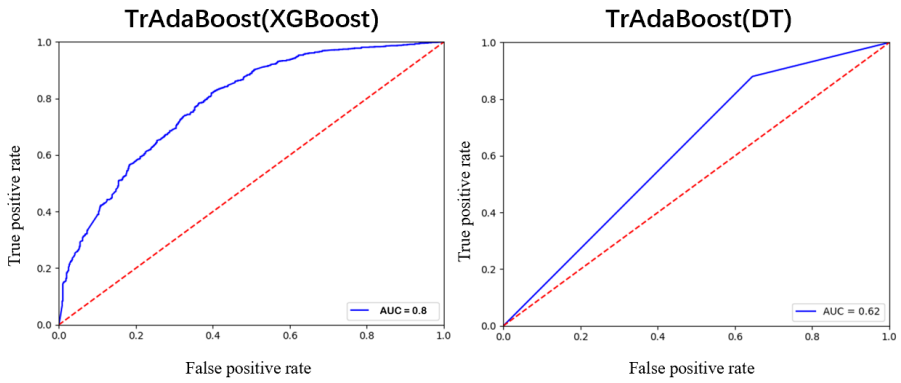

Fig. 4. Comparison results-1

It can be seen from the experimental results that using the Ensemble Learning algorithm XGBoost as the Base Learner increases the AUC value of the base learner by 18% compared with the simple algorithm decision tree as the Base Learner. Therefore, it reveals that the choice of Base Learner has an important influence on the final result.

To demonstrates the superiority of transfer learning algorithm, we also select decision tree, XGBoost, Logistic regression algorithm to predict the target domain respectively. Observe the results of training using only the target domain data and the models in this paper. The results are shown in table 6 and Figure 5.

Table 6. Comparison results-2.

TL VS ML	TrAdaBoost（XGBoost）	XGBoost	Decision Tree	Logistic
AUC	0.80	0.56	0.61	0.64
prediction	0.79	0.59	0.61	0.62
recall	0.65	0.59	0.64	0.67
F1	0.71	0.59	0.61	0.64

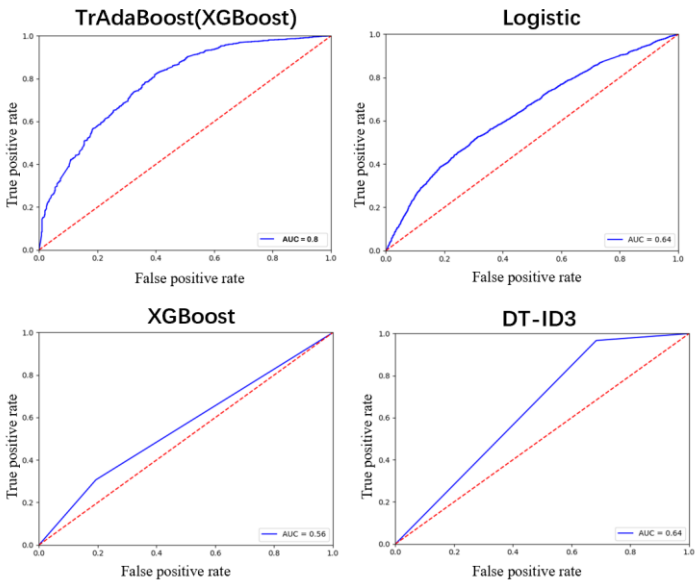

Fig. 5. Comparison results-2

From this, it is clear that using transfer learning algorithms to train the target domain has a higher AUC, prediction, recall and F1 than traditional machine learning. It also further verifies that transfer learning algorithms can better solve the prediction of small samples problem.

5 Conclusion

We constructs person Credit Evaluating Model based on Instance-based Transfer Learning, and focuses on the choice of Basic Learners in the design. The model shows better classification and forecasting capabilities, and can help banking and P2P financial institutions to avoid risks to a certain extent. Besides, the model uses Information divergence to select features with greater contribution to reduce computational complexity. We do a lot of experiments to select the Base Learner and improve the accuracy of the model. The TrAdaBoost (XGBoost) model makes full use of the source domain information to successfully complete the training of the target domain information, and solves the predicament that the data set cannot be trained due to the lack of samples and significant missing values. This article achieves the transfer of samples in the field of personal credit risk, which has certain reference value for the financial field. The model based on TrAdaBoost sample transfer proposed in this paper adds the XGBoost Ensemble Learning algorithm, which improves the accuracy of the model, enhances the performance of the model, and has good generalization capabilities.

References

1. Shan He, Zhendong Liu, Xiaolin Ma. A comparative review of credit scoring models——Comparison between traditional methods and data mining ,CREDIT REFERENCE, 2019, 037(002):57-61.
2. Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. IEEE TKDE, 22(10):1345–1359.
3. Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. Journal of Big Data, 3(1):1–40.
4. Dai, W., Yang, Q., Xue, G.-R., and Yu, Y. (2007). Boosting for transfer learning. In ICML, pages 193–200. ACM
5. Cook et al., 2013] Cook, D., Feuz, K. D., and Krishnan, N. C. (2013). Transfer learning for activity recognition: A survey. Knowledge and information systems, 36(3):537–556）
6. Kermany et al., 2018] Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., McKeown, .A., Yang, G., Wu, X., Yan, F., et al. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell, 172(5):1122–1131.
7. ZHUANG Fu-Zhen, LUO Ping, HE Qing, SHI Zhong-Zhi. Survey on Transfer Learning Research[J]. Ruan Jian Xue Bao/ Journal of Software, 2015, 26(1): 26-39.http://www.jos.org.cn/1000-9825/4631.html
8. Xiulong Han. “User credit rating modeling based on xgboost” Computer Knowledge and Technology 5(2018).
9. Zhu Y, Chen Y, Lu Z, Pan SJ, Xue GR, Yu Y, Yang Q. Heterogeneous transfer learning for image classification. In: Burgard W,Roth D, eds. Proc. of the AAAI. AAAI Press, 2011. 1304-1309.
10. Pan W, Xiang EW, Yang Q. Transfer learning in collaborative filtering with uncertain ratings. In: Hoffmann J, Selman B, eds. Proc.of the AAAI. AAAI Press, 2012. 662-668.
11. ZHU Bing，HE Chang-zheng，LI Hui-yuan. Research on Credit Scoring Model Based on Transfer Learning,OPERATIONS RESEARCH AND MANAGEMENT SCIENCE, 24.002(2015):201-207.
12. Zheng, Lutao , et al. "Improved TrAdaBoost and Its Application to Transaction Fraud Detection." IEEE Transactions on Computational Social Systems PP.99(2020):1-13.
13.]Xu Wang. Application of transfer learning in quantitative stock selection. Diss. 2019
14. Zhongzhong Yu.”Ensemble transfer learning algorithm for imbalanced sample classification” Acta Electronica Sinica40.007(2012):1358-1363.
15. Dai, W., Yang, Q., Xue, G.-R., and Yu, Y. (2007). Boosting for transfer learning. In ICML, pages 193–200. ACM
16. Chen, Tianqi , and C. Guestrin . "XGBoost: A Scalable Tree Boosting System." (2016).
17. LIU Jian cheng，JIANG Xm hua，WU Jm pei,Realization of a Knowledge Inference Rule Induction System[J],Systems Engineering,2003, 21(3): 108-110.