SVM Model Against Telecom Card Fraud Using GA Optimised Ten-Fold Cross-Testing

Peng Wei *

School of International College of Transportation, Chang'an University, Xi'an, China, 710000

* Corresponding Author Email: peng.wei@ucdconnect.ie

Abstract. The increased number of payment methods also makes it easier for personal information to be stolen by criminals, and for criminals to take over financial payment accounts and steal money. With trillions of bank card transactions occurring every day, Credit Card Fraud Detection (CCFD) is a serious challenge, so this paper predicts "whether or not fraud occurs" by using six types of machine learning models. For problem 1, firstly, "mean, maximum, minimum, median, variance, standard deviation, quartile" are calculated for each indicator; secondly, data cleaning is carried out, and the data set is found to be free of missing values and outliers. Then the data preprocessing work was carried out, min_max normalisation and z-score standardisation were performed on the data. After that, correlation analysis was carried out, and the first four indicators were classified as negative indicators and the last three as positive indicators according to the characteristics of the indicators themselves. It can be found by calculating the Pearson correlation coefficient value after two data processing. Using the coefficient of variation method to calculate the weight of the seven "influence whether fraud" indicators. Finally, BP neural network model, decision tree model, random forest classification model, ELM model, SVM model, logistic regression model are established. For Problem 2, the four models constructed in Problem 1 are solved; to solve the BP neural network model: the data set is divided into training set and testing set according to the ratio of 6:4, and the sigmod function is used as the activation function. For BP neural network, "output >0.5" is recorded as 1, i.e. fraudulent behaviour; "output <0.5" is recorded as 0, i.e. non-fraudulent behaviour. Adjusting the learning rate and the number of iterations, the optimal average mean square error after optimal gradient descent is smaller. To solve the SVM model, the data set is divided into ten groups using the improved ten-fold cross-test, with one group as the training set and nine groups as the validation set, so as to obtain the model with the highest accuracy and the corresponding training data, and then the genetic algorithm is used to search for the optimisation of the kernel parameters in the SVM model on this basis. To solve the decision tree model, the training set and prediction set are divided into 7:3 and solved, and the number of leaf nodes is optimised. Solve the random forest classification model, divided into training set and prediction set according to 7:3 and solved, for similar accuracy choose the random forest classifier when the decision tree is less.

Keywords: Machine learning, correlation analysis, data normalisation, genetic algorithms.

1. Introduction

At present, with the popularization of credit cards, credit card fraud increases gradually. Digital payment mainly refers to the digital payment methods achieved with the help of hardware facilities such as computers and smart devices and digital technology such as communication technology, artificial intelligence and information security. Currently, globally, it is increasing from 35% in 2014 to 57% in 2021 in developing economies [1,2].

According to the Data Breach Index (IBM Annual Report on Information Breaches), more than 5 million records are stolen every day, a statistic that demonstrates that fraud is still very prevalent for both carded and cardless types of payments. The reasons for this are closely related to both the current regular features such as the full spread of crime to the Internet. Existing methods have a number of serious drawbacks in practice, such as severely diminished utility under privacy restrictions or an excessive amount of communication between the information fusion center and local data providers. [3]. Misclassifications impact today's fraud detection systems, and high false positive rates limit their utility [4]. Wiese et. al. addresses this problem by proposing the use of a dynamic machine learning method in an attempt to model the time series inherent in sequences of same card transactions [4].

On the basis of two-factor authentication techniques, the goal is to provide methods and algorithms for integrity control and authenticity of data packets in banking transaction security protocols. [5]. On a dataset of nearly 80 million credit card transactions that have been pre-labeled as fraudulent and legitimate, evaluate a subsection of Deep Learning topologies, from the general artificial neural network to topologies with built-in time and memory components such as Long Short-term memory — and different parameters with regard to their efficacy in fraud detection [6]. In order to address the unbalanced data distribution, Hassan et. al. suggests a cutting-edge insurance fraud detection technique [7]. Based on the above background, this study intends to address the following questions.

(1) Downsampling and data cleaning of the data first, and then using data normalisation to limit the range of features to avoid different features being mistaken by the algorithm as being more important due to large differences in values. After normalization, the data will be stripped of the original dimension and distributed in the range of 0 to 1 [8]. A variety of "machine learning models" for data mining are used. The ultimate goal of this paper is to identify models with better classification and prediction ability to analyse credit card transaction data, and to improve the accuracy and efficiency of prediction as much as possible without causing overfitting.

(2) The given sample data are further mined and analysed, and the parameters of different models are adjusted through cross-validation, mesh tuning, optimisation of leaf nodes, adjustment of learning rate, decision tree pruning and different activation functions, to find the optimal solution, and then several optimal models are compared.

2. Problem analysis

2.1. Analysis of question one

Statistical analyses and calculations were performed for seven indicators that affect fraud results (distance from home of the card transaction location, distance from the last transaction, ratio of the most recent transaction to the median price of previous transactions, whether the transaction took place at the same merchant, whether the transaction was made through a chip, whether a PIN code was used for the transaction, and whether the transaction was an online order). For each indicator, "mean, maximum, minimum, median, variance, standard deviation, and quartile" are calculated, so that they can be used to deal with outliers, missing values, and duplicated values in data cleaning. Thereafter, data cleaning is performed to determine whether there are missing values, outliers, and duplicate values. If there are outliers and missing values, they should be processed by deleting and interpolating them. And the data preprocessing of the cleaned data; this time, the "min_max normalisation" and "z-score normalisation" two normalisation methods are used, and the z-score normalisation converts the data of different scales into the same scale. The z-score normalisation converts the data of different scales into the same scale. Normalisation can accelerate the speed of finding the optimal solution in gradient descent; however, it is also necessary to keep the "un-normalised" data, which is used in part of the model. After pre-processing, each indicator is analysed for correlation and the correlation coefficient is calculated. The correlation coefficient is calculated to determine the influence of each indicator on "whether fraud occurs". Finally, the coefficient of variation method was used to calculate the weights of the seven indicators.

2.2. Analysis of question two

To solve the BP neural network model, For BP neural network, the dataset is divided into training set and test set in the ratio of 6:4 and sigmod function is applied as activation function. For BP neural network output >0.5 is recorded as 1(fraudulent behaviour); output <0.5 is recorded as 0.

To solve the SVM model, considering the large amount of data, the SVM computing rate is slow, so the improved ten-fold cross-checking method is used, the data set is divided into ten groups, one group for the training set, and the other nine groups for the validation set, so as to obtain the model with the highest accuracy and the corresponding training data, based on which genetic algorithms are used to search for the optimality of the kernel parameter in the SVM, in order to achieve further

optimization of the model, improve the accuracy. SVM model is optimized by genetic algorithm, compatible and efficient global search ability can reduce the influence of key parameter setting on evaluation results to a certain extent [9].

To solve the decision tree model, The data set is divided into training set and prediction set in the ratio of 7:3 for solving, and then the number of leaf nodes is optimised, and the optimal number of leaf nodes and the highest accuracy rate are obtained by iterating the accuracy rate under different numbers of leaf nodes, on the basis of which, pruning of the decision tree reduces the complexity of the model on the basis of guaranteeing the accuracy rate, prevents the model from overfitting, and improves the robustness and stability of the model.

To solve the random forest model, the dataset is divided into training set and prediction set according to the ratio of 7:3, and the decision tree with the highest accuracy rate of random forest classification is searched by iterating the "number of decision trees". For random forest, the number of required samples is small, the classification speed is fast, and the prediction accuracy can be significantly improved [10]. The random forest classifier with fewer decision trees is selected when the accuracy rate is very different, so as to prevent the model from being overfitted and to make the model's generalisation The performance of Random Forest Classifier is better.

3. Data collection and pre-processing

3.1. Statistical analysis calculations

The seven impact indicators are labelled as ①, ②, ③, ④, ⑤, ⑥, and ⑦ for the purpose of description, as shown in Table 1.

Table 1. Relationship between indicator number and corresponding indicator

Num.	Corresponding indicators
①	Distance between card transaction location and home
②	Distance from the occurrence of the last transaction
③	Ratio of recent transaction to median price of previous transactions
④	Whether the transaction took place at the same merchant
⑤	Whether or not the transaction is made through a chip (bank card)
⑥	Whether a PIN code was used for the transaction
⑦	Whether it is an online trading order

Statistical calculations were made on the sample data for each indicator and the results are shown in Table 2 below:

Table 2. Statistical calculations

	①	②	③	④	⑤	⑥	⑦	Fraud
summation	26628792	5036519	1824182	881536	350399	100608	650552	87403
mean	26.62879	5.036519	1.824182	0.881536	0.350399	0.100608	0.650552	0.087403
maximum	10632.72	11851.1	267.8029	1	1	1	1	1
minimum	0.004874	0.000118	0.004399	0	0	0	0	0
median	9.96776	0.99865	0.997717	1	0	0	1	0
variance	4275.955	667.8655	7.837698	0.10443	0.22762	0.090486	0.227334	0.079764
standard deviation	65.39078	25.84309	2.799589	0.323157	0.477095	0.300809	0.476796	0.282425
quartile	3.878008	0.296671	0.475673	1	0	0	0	0

3.2. Correlation analysis

In order to find the hidden correlation of things statistics accurately, the qualitative analysis of data is done first, and then the quantitative relationship of things is analyzed [11]. Determine the positive indicators, negative indicators:

The 'distance between the bank card transaction location and the home', 'the distance from the last transaction', 'the ratio of the last transaction to the median price of the previous transaction', and 'whether it is an online transaction order' are the first four positive indicators.

The last three indicators, 'whether the transaction occurs in the same merchant', 'whether the transaction is performed through the chip', and 'whether the PIN code is used during the transaction', are used as negative indicators. The z-score normalized data was subjected to ' correlation coefficient calculation ', and the results are shown in Table 3.

Table 3. Pearson correlation coefficient values solved after z-score standardization.

	①	②	③	④	⑤	⑥	⑦	fraud
①	1	0.000193	-0.00137	0.143124	-0.0007	-0.00162	-0.0013	0.187571
②	0.000193	1	0.001013	-0.00093	0.002055	-0.0009	0.000141	0.091917
③	-0.00137	0.001013	1	0.001374	0.000587	0.000942	-0.00033	0.462305
④	0.143124	-0.00093	0.001374	1	-0.00134	-0.00042	-0.00053	-0.00136
⑤	-0.0007	0.002055	0.000587	-0.00134	1	-0.00139	-0.00022	-0.06097
⑥	-0.00162	-0.0009	0.000942	-0.00042	-0.00139	1	-0.00029	-0.10029
⑦	-0.0013	0.000141	-0.00033	-0.00053	-0.00022	-0.00029	1	0.191973
fraud	0.187571	0.091917	0.462305	-0.00136	-0.06097	-0.10029	0.191973	1

3.3. Coefficient of variation method to calculate the index weight.

The results of calculating the index weight by the coefficient of variation method are shown in table 4. After solving the weight value by the coefficient of variation method, the weight of No.6 index (whether PIN code is used during the transaction) is the largest, which is 55.7114 %. The second is the No.5 indicator (the transaction through the chip), which has a larger weight of 25.37 %. The weight of No.2 index (the distance from the last transaction) is the smallest, which is 0.1801 %.

Table.4. Pearson correlation coefficient values solved after z-score standardization.

Index	①	②	③	④	⑤	⑥	⑦
Weight ω_i	0.001046	0.000349	0.001801	0.049418	0.253704	0.557114	0.136564

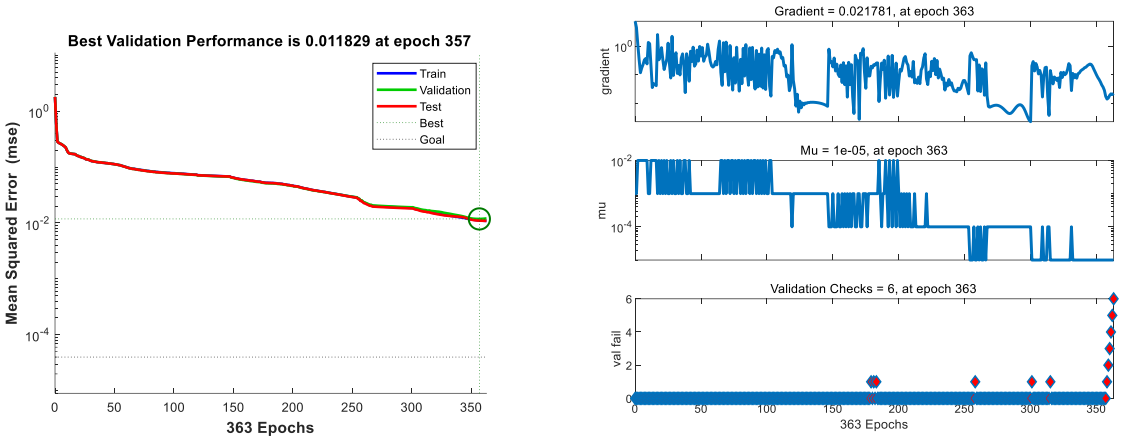
4. Model construction and analysis

4.1. Solving the BP neural network model.

Bhatia et. al. examines SVM and neural networks for the detection of credit card fraud [12].

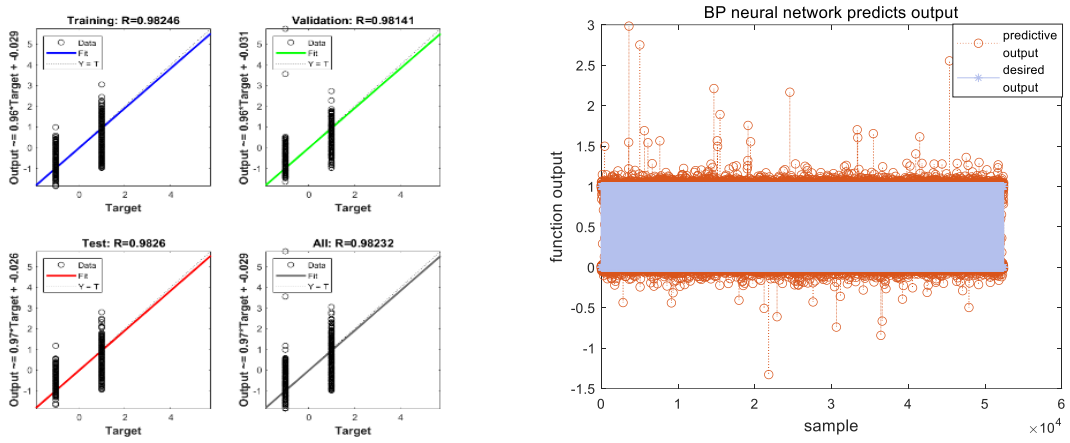
In order to avoid the contingency caused by the data set arrangement, all data are randomly sorted to avoid statistical system errors. The data set is segmented and labeled, and a single variable is pursued as much as possible to compare the optimal prediction algorithm for this scenario under the premise of sufficient computational efficiency.

For BP neural network, the data set is divided into training set and test set according to the ratio of 6: 4, and the sigmod function is used as the activation function. For BP neural network output > 0.5 is recorded as 1, that is, fraud; output < 0.5 is recorded as 0, that is, non-fraud behavior. Constantly adjust the learning rate and the number of iterations to find the optimal gradient descent results (Figures 1 and 2). It can be seen from Figure 1 (a) that the best performance of the BP neural network is the 522nd round, and the mean square error is small, which is 0.17094.



(a) (b)

Figure. 1 The relationship between the best training performance and the mean square error (a) and the trend of gradient descent, validation set and learning rate (b)



(a) (b)

Figure. 2 The relationship between the fitting image (a) and the predicted output and the expected output (b) under different R values.

4.2. Solve the SVM model.

4.2.1 Using ten-fold cross validation.

Cross Validation, sometimes called Rotation Estimation, is a statistically practical method of cutting data samples into smaller subsets, a theory proposed by Seymour Geisser.

In a given modelling sample, take out most of the samples for modelling, leave a small portion of the samples to be forecast with the model just built, and find the forecast error for this small portion of the samples and record their sum of squares. This process continues until all samples have been forecast once and only once. Summing the squares of the forecast errors for each sample is called PRESS (Predicted Error Sum of Squares).

Faced with a massive amount of data of 1 million entries, it is difficult for individual user computers to ensure speed and efficiency while running the complete computation at the same time. For this reason, this study adopts an improved S-fold cross-checking approach, which treats the data set as a 10-fold cross-validation system, which means that the training and testing algorithms will be run iteratively for 10 times, the first time slicing 1-100,000 in the document for training and dividing the remaining 900,000 pieces of data into 9 groups to test them on the training model, and so on. Generally cross-checking divides, the data into 10 groups, but due to the sheer volume of data, the actual saving in computational power is only 10%. For this reason, this study calculates the average accuracy of each model and compares them after the cross-testing operation and selects the model with the highest accuracy as the best-fitting model.

4.2.2 Using Genetic Algorithm for Optimal Models

In order to develop and train numerous fuzzy rule-based classifiers (FRBCs) to recognize patterns of financial statement fraud, (Alden et. al., 2012) uses a genetic algorithm (GA) [13]. Saheed et. al. concentrate on the feature selection process for CCF detection using Genetic Algorithm (GA). [14]. The author uses a genetic algorithm to encode the solution of the problem as a chromosome and screen out individuals with high fitness values according to the probability distribution of the fitness function. Then the most viable chromosomes are made to survive with the highest probability by the three basic genetic operators of selection, crossover and mutation, and the population evolves to better and better regions in the search space generation by generation. Finally, the optimal individuals in the last generation of the population can be decoded as the near-optimal solution that meets the optimization objective.

4.2.3 Solution results

The improved ten-fold cross-checking method is used, and the data set is evenly divided into ten groups, one group is used as the training set, and the other nine groups are used as the validation set, so that the model with the highest accuracy and the corresponding training data can be obtained, based on which genetic algorithms are used to find the optimal kernel parameter c and g in the SVM, and the accuracy of the optimised genetic algorithms is 96.35%, with the optimal parameter c = 84.6476 and the optimal parameter g = 0.0035286. The horizontal coordinate is the number of evolutionary generations, and the vertical coordinate is the degree of adaptation, with a starting number of generations of 0 and a terminating number of generations of 100, gradually increasing from 0 to 100 at intervals of 10, with average and optimal adaptations above 90% (Figure 3).

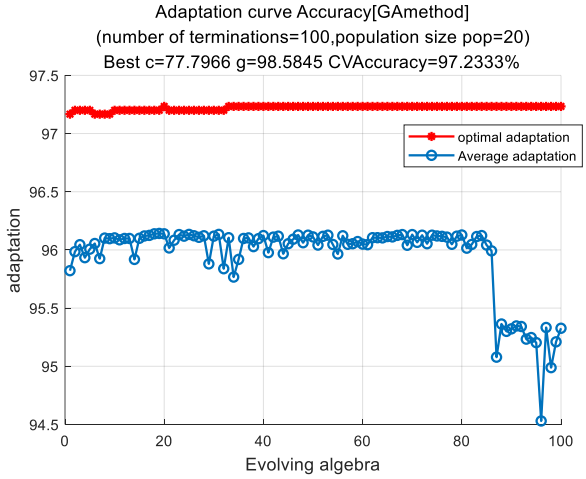
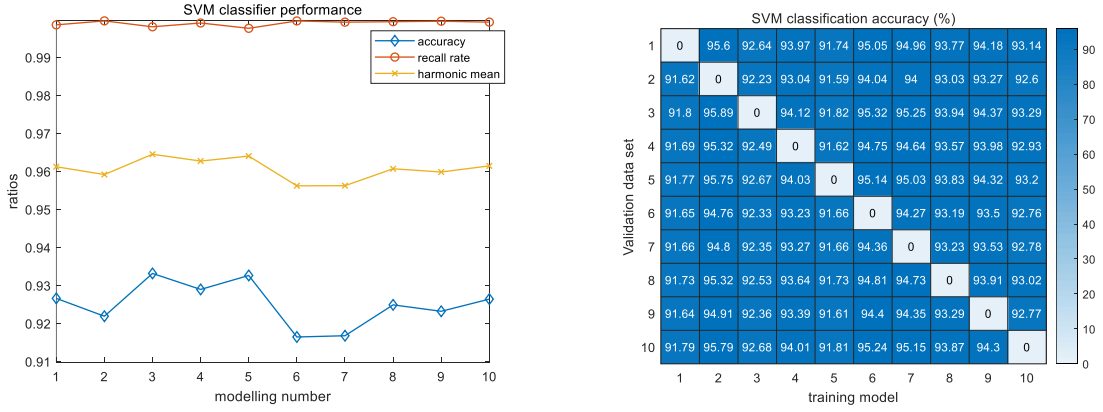


Figure 3. Trends in the number of generations of evolution versus fitness



(a) (b)

Figure 4. Trends in precision, recall, and reconciliation mean (a) and SVM model heat map (b)

The rates of precision, recall, and reconciled means for each model are all above 90%, with the highest recall and a late recall of one.

The data is divided into 10 parts, each data corresponds to a model, the horizontal coordinate is the number of 10 models, the vertical coordinate is the number of 10 sets of data sets, and the number in each block in the Figure 4 indicates the classification accuracy of the corresponding model and the corresponding data.

The accuracy of the optimised model through genetic algorithm optimisation reaches up to 96.35%, as shown in Figure 5, and the actual test set classification is close to the predicted test set classification.

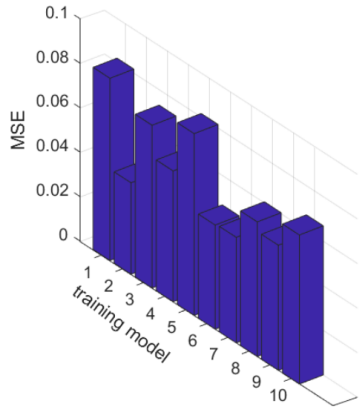


Figure 5. Relationship between the training model and the average put error

4.3. Solving the decision tree model

To improve the accuracy of fraud detection, Prusti et. al. suggests using widely used classification algorithms like Decision Tree (DT) and Support Vector Machine (SVM) [15]. Decision tree, the dataset is divided into training set and prediction set in the ratio of 7:3 for solving, then the number of leaf nodes is optimised, and the optimal number of leaf nodes and the highest accuracy rate are obtained by iterating the accuracy rate under different number of leaf nodes, based on which, the complexity of the model is reduced on the basis of guaranteeing the accuracy rate, preventing the model from overfitting, and improving the model's robustness and stability(Figure 6).

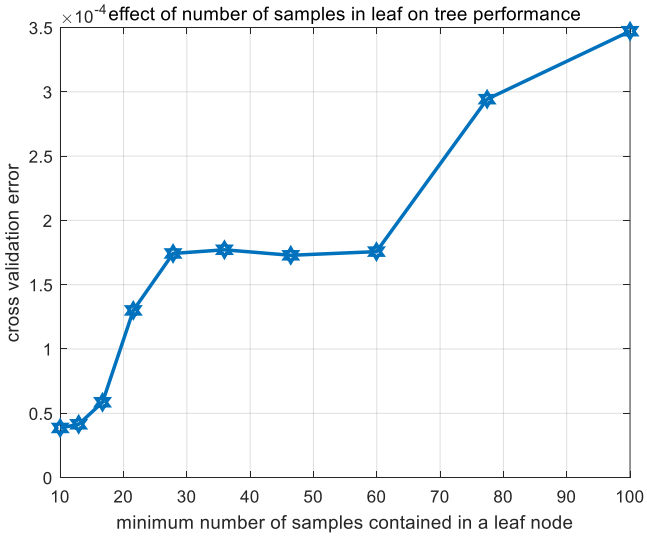


Figure 6. Effect of minimum number of samples contained in leaf nodes on the performance of decision tree.

After setting the minimum number of samples contained in the leaf nodes to 12, the optimised decision tree is generated, and its performance is compared with the original decision tree performance.

The optimised decision tree classifier is much simpler, although the re-substitution error of the optimised decision tree classifier is higher than that of the pre-optimised decision tree classifier, the cross-validation error is the same.

The resampling error after pruning is 0, which is good.

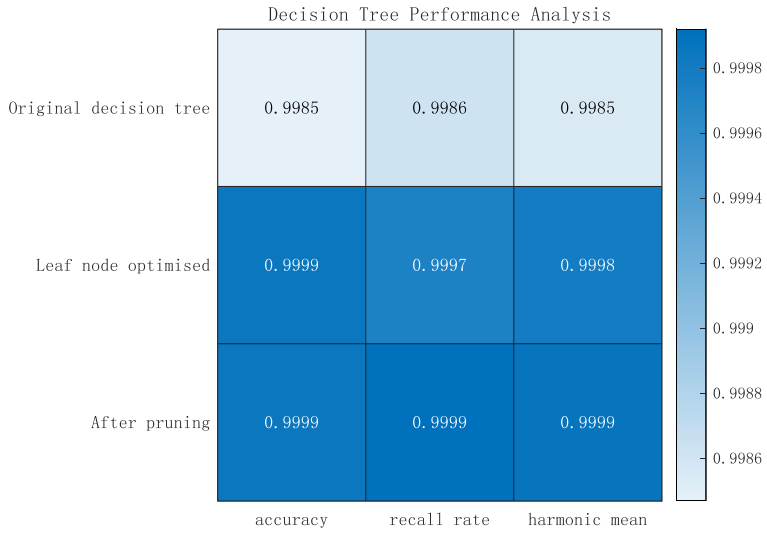
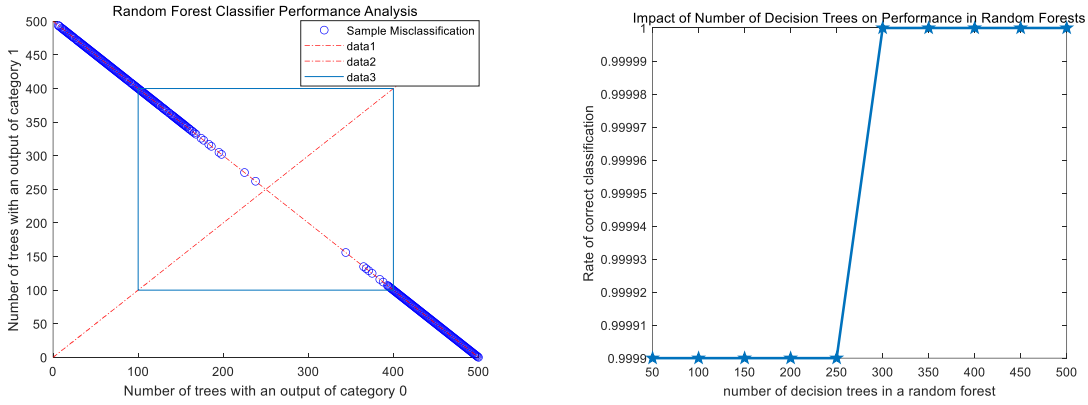


Figure 7. Decision tree model heat map

The heat map depicts the magnitude and change of precision, recall, and reconciliation mean of the original decision tree, leaf node optimisation, and pruning (Figure 7).

4.4. Solving Random Forest Classification Model

The dataset is divided into training set and prediction set according to the ratio of 7:3 for solving, and the number of decision trees is iterated to find the decision tree with the highest accuracy rate of Random Forest Classification, and for the difference in the accuracy rate is very small, the Random Forest Classifier with fewer decision trees is selected, so as to prevent the model from being overfitted, and to make it have a better generalisation performance.



(a) (b)

Figure 8. Distribution of locations of misclassified samples (a) and the effect of decision tree countability on classification correctness in random forests (b)

The horizontal coordinate is the decision tree with output category 0, the vertical coordinate is the decision tree with output category 1, and the blue colour refers to misclassified samples (Figure 8).

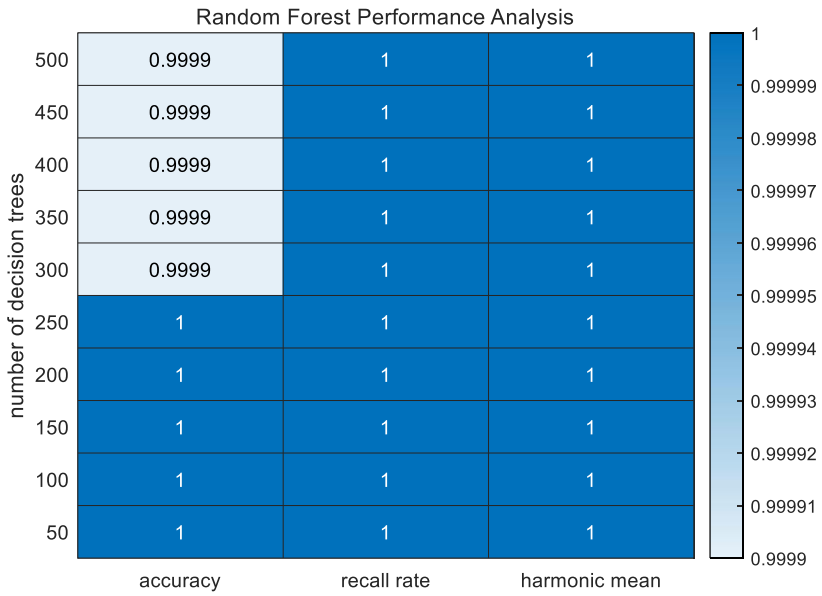


Figure 9. Random Forest classification model heat map

Different number of decision trees corresponds to different precision, recall and F1, with the increase of the number of decision trees, the precision rate changes from 1 to 0.9999, the recall rate, and the mean of reconciliation is always 1(Figure 9).

5. Conclusions

In order to study the prediction of telecommunication fraud results, downsampling is used to balance the data size and avoid overfitting, and the feature values are normalised so that the absolute size of the system values becomes relative size to avoid the algorithm to misidentify that the features with large values such as distance are more important than the percentage. For data prediction this study establishes BP neural network, SVM, decision tree and random forest model. The use of pruning to optimise the decision tree model, the final results to see the tree model and random forest model is most suitable for data classification and prediction, the results can be rounded up to close to 100%, the model can guide banks and other financial institutions to remittances involved in the analysis of the transaction, abnormal transfers to remind the transaction risk, confirm the purpose of the transaction or to freeze the transfer of funds in order to protect the user's property security. In addition, this paper uses the genetic algorithm to optimise the neural network model and SVM model can effectively solve the problem of model parameter assignment, with the advantage of the genetic algorithm's global search and efficient parallelism to obtain and accumulate the knowledge of the search space and adaptively control the process in order to obtain the optimal solution, and for the SVM model, the results of the study in the prediction of the bank card land test also has a certain degree of reliability, can be used for more similar models. This model can be used in more similar telecommunication fraud incidents.

References

[1] Wang C, Han D. Credit card fraud forecasting model based on clustering analysis and integrated support vector machine [J]. Cluster Computing, 2019, 22: 13861-13866.

[2] Zhao Qi. Global digital payment is growing [N]. Chinese Journal of Social Science,2022-07-08(003).

[3] Wang Y, Adams S, Beling P, et al. Privacy preserving distributed deep learning and its application in credit card fraud detection [C]//2018 17th IEEE International Conference on Trust, Security and Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). IEEE, 2018: 1070-1078.

[4] Wiese B, Omlin C. Credit card transactions, fraud detection, and machine learning: Modelling time with LSTM recurrent neural networks [M]//Innovations in neural information paradigms and applications. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009: 231-268.

[5] Evseev S P, Tomashevskyy B P. Two-factor authentication methods threats analysis [J]. Радіоелектроніка, інформатика, управління, 2015 (1 (32)): 52-59.

[6] Roy A, Sun J, Mahoney R, et al. Deep learning detecting fraud in credit card transactions [C]//2018 systems and information engineering design symposium (SIEDS). IEEE, 2018: 129-134.

[7] Hassan A K I, Abraham A. Modeling insurance fraud detection using imbalanced data classification[C]//Advances in Nature and Biologically Inspired Computing: Proceedings of the 7th World Congress on Nature and Biologically Inspired Computing (NaBIC2015) in Pietermaritzburg, South Africa, held December 01-03, 2015. Springer International Publishing, 2016: 117-127.

[8] Zhang Ge. Research on data preprocessing in Course Recommendation Prediction Model [J]. China New Communications,2019,21(19):185.

[9] Yang Kang, XUE Xicheng, LI Shibo. Geological hazard susceptibility evaluation based on GA-optimized SVM model with information integration [J]. Safety and Environmental Engineering,2022,29(03):109-118.

[10] Xu Ge, Zhang Ke. Real Estate Price Evaluation based on Random Forest Model. Statistics and Decision,2014(17):22-25.

[11] Lin Tingting. Research on Grade Prediction Model based on BP Neural Network Algorithm [J]. Computational Technology and Automation,202,41(01):79-81+147.

[12] Tyagi C S, Parwekar P, Singh P, et al. Analysis of Credit Card Fraud Detection Techniques [J]. Solid State Technology, 2020, 63(6): 18057-18069.

[13] Alden M E, Bryan D M, Lessley B J, et al. Detection of financial statement fraud using evolutionary algorithms[J]. Journal of Emerging Technologies in Accounting, 2012, 9(1): 71-94.

[14] Saheed Y K, Hambali M A, Arowolo M O, et al. Application of GA feature selection on Naive Bayes, random forest and SVM for credit card fraud detection[C]//2020 international conference on decision aid sciences and application (DASA). IEEE, 2020: 1091-1097.

[15] Prusti D, Rath S K. Web service-based credit card fraud detection by applying machine learning techniques[C]//TENCON 2019-2019 IEEE Region 10 Conference (TENCON). IEEE, 2019: 492-497.