Discrimination of Insurance Fraud Based on Machine Learning

Tianqi Yang *, Yue Wu

Faculty of Business and Management, Beijing Normal University- Hongkong Baptist University United International College, Zhuhai, Guangdong, China

* Corresponding Author Email: tqyoung@126.com

Abstract. In the insurance industry, Insurance fraud is a common phenomenon. However, according to statistics, among all types of insurance, automobile insurance fraud is the high incidence time of Insurance fraud. Based on 39 characteristic variables in the insurance claims database, this paper completes data preprocessing through normalization and coding of Categorical variable; Then it analyzes the correlation between data characteristics and Insurance fraud; Then, the principal component analysis method is used to reduce dimensionality and extract features from multi-dimensional features; Finally, the SVM classifier is trained to effectively identify Insurance fraud. The research results show that the model is effective in identifying Insurance fraud, and can achieve 79% accurate discrimination.

Keywords: Insurance Fraud, Normalization, PCA, SVM.

1. Introduction

With the continuous improvement of China's economic level, the people's highest demand for living standards is increasing, and the number of domestic motor vehicles is also constantly increasing. As an important source of income for Property insurance companies, motor vehicle insurance business accounts for more than 50% of the Property insurance market. According to the statistics, about 20% of the automobile insurance claims are likely to have Insurance fraud, but less than 3% of the fraud can be identified, and fraud is becoming increasingly serious. In recent years, the comprehensive reform of automobile insurance has also put forward the goal of "reducing prices, increasing insurance coverage and improving quality", which conforms to the policy of our country. Through efficient Insurance fraud discrimination, insurance funds can be kept in good operation and maintain good social fashion.

With the rapid development of Big data, artificial intelligence and other technologies, the financial supervision field has begun to shift from traditional artificial supervision to intelligent Big data supervision [1-3]. Through the idea of data mining, intelligent supervision helps to process Insurance fraud data with Big data and multi feature dimensions, and excavates potential Insurance fraud behaviors [4-7]. Intelligent Insurance fraud discrimination is a machine learning, deep learning and other data Discriminative model with the ability of Self Supervised Learning [8-11]. It can achieve the accuracy of fraud recognition according to the characteristics of high correlation with Insurance fraud, and then adapt to the emergence of the changing new model of Insurance fraud, so as to achieve more accurate discrimination and curb the occurrence of Insurance fraud.

In this paper, theoretical analysis and empirical analysis are combined to build a model. In terms of data processing and feature extraction, principal component analysis is used. In the training process of the classifier, support vector machine and linear kernel function are used to realize accurate discrimination of Insurance fraud. The above are the characteristics of the model in this paper.

2. Data preprocessing

Data from website https://tianchi.aliyun.com. This data set provides the data of automobile insurance claims from customers, including 700 samples, each of which has 39 variables. The last one represents whether the customer has Insurance fraud, and the remaining 38 are the customer's background information or accident information.

2.1. Data processing

To distinguish Insurance fraud, it is necessary to select all variables related to Insurance fraud, and eliminate variables that have little or no relevance to Insurance fraud. This paper excludes two sample coding variables: policy id (insurance number) and insured zip (postal code of the insured). These two variables have obvious weak correlation with Insurance fraud.

The three data in the database, Incident date, policy bind date, and auto year, are in the form of dates and cannot be directly involved in calculations. Considering practical processing, this article redefines two time indicators: delta time1=time difference from insurance binding to accident occurrence=accident date - insurance binding date, Delta time2=time difference from car purchase to accident occurrence=accident date - car purchase date, and delete the original 3 date indicators. The data types of delta time1 and delta time2 are real numbers.

2.2. Coding of data categorical variable

In the dataset, sample variables are divided into numerical variables and Categorical variable. Numeric variables are very meaningful for operations such as addition, subtraction, and averaging. Database digital variables include customer months, total claims, income claims, property claims, vehicle claims, etc. Categorical variable is a name that describes the category of things. They are meaningless for operations such as addition, subtraction, and averaging. The database Categorical variable include single limit of policy portfolio, insured gender, automobile manufacturer, Car model, etc.

Therefore, this article does not deal with numerical variables. For the Categorical variable, in order to prevent over fitting, the label encoding method is used, which means encoding with labels, that is, encoding the original eigenvalue into a user-defined digital label to complete the quantization encoding process. Label encoding solves the problem of classification encoding and allows for the freedom to define quantified numbers.

3. Methodology

As the number of Insurance fraud is increasing, in many research methods, this paper uses machine learning to accurately identify Insurance fraud. The specific steps are as follows:

1) Firstly, normalize the variables of the sample to eliminate the adverse effects caused by the singularity of the sample data;

2) Data correlation analysis to obtain the correlation between various variables in the sample database and Insurance fraud, and extract the attributes with high sample correlation for analysis;

3) Principal component analysis is carried out on the attributes with high sample correlation to obtain the characterization with the highest correlation to Insurance fraud;

4) The support vector machine is used to classify and train the sample database to realize accurate identification of Insurance fraud.

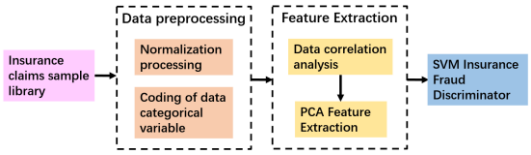

Figure 1. Design flowchart of insurance fraud discriminator

3.1. Data normalization processing

The insurance claims dataset has 700 samples, each with 39 attribute variables. For each variable, the range of numerical changes is different, such as total_Claim_The total claim amount is within the range of ten thousand yuan (RMB) and the age is within the range of ten. Therefore, in the insurance claim dataset, the evaluation of different indicators (i.e., different features in the feature vector are different evaluation indicators) often have different dimensions and dimensional units, which will

affect the results of data analysis. Therefore, in order to eliminate the dimensional impact between indicators, data standardization is necessary to solve the comparability between data indicators. After normalization processing of the original data, all indicators are in the same order of magnitude, which can accelerate the speed of gradient descent to find the optimal solution and improve the accuracy of discrimination.

The linear function converts the linearization method of the original data to the range of [0, 1]. The normalization formula is as follows:

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Where x^* is the normalized data, x is the original data, and x_{max} and x_{min} are the maximum and minimum values of the original dataset, respectively.

3.2. Data correlation analysis

There are 39 attribute variables in each sample of insurance claim data set, and these attribute variables have strong influence terms and if influence terms for the discrimination of Insurance fraud. In order to reduce the interference of the influencing variables, this paper conducts correlation analysis on 38 attribute variables in the insurance claim data set through Pearson correlation coefficient to extract attribute variables that have strong correlation with Insurance fraud.

The Pearson Correlation Coefficient (PCC) used in this article is a method for calculating linear correlation. The Pearson correlation coefficient is the ratio of covariance to standard deviation.

The Pearson correlation coefficient between two variables in the sample is obtained by the quotient of covariance and standard deviation between the two variables:

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y}, \tag{2}$$

Where x and y are two random variables, and $cov(x, y)$, which are the covariance of x and y, σ_x is the standard deviation of x, σ_y is the standard deviation of y.

For database samples, formula (2) can be rewritten as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{3}$$

Where x and y are two random variables, and n is the number of samples, x_i, y_i are the i-point observations corresponding to variables x, y, $\bar{x}$ is the average of x samples, $\bar{y}$ is the average of y samples.

The correlation calculation was performed on 39 variables of 700 samples in the insurance claims database, and the results are shown in Figure 2.

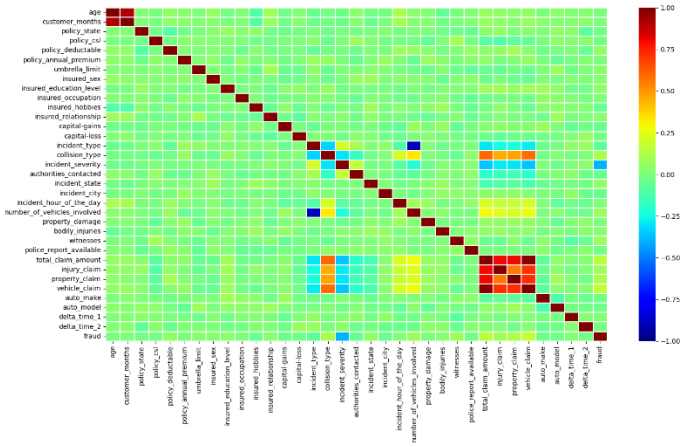

Figure 2. Correlation matrix based on Label encoding

3.3. Feature Extraction

Through the above correlation proof, this paper has carried out strong correlation feature extraction for 38 sample variables in the sample database, but for the Discriminative model, it still has too many data dimensions, high data dimensions. Due to the sparsity of data, the comparability of distance between data objects will no longer exist, and its effectiveness will be greatly reduced; At the same time, encountering dimensional expansion will exponentially increase the computational workload. Therefore, this article adopts PCA dimensionality reduction to solve high-dimensional problems.

The PCA algorithm is a method specifically used to reduce the dimensionality of high-dimensional data. By reducing the dimensionality of high-dimensional data to obtain low dimensionality, the training speed of the model can be accelerated, and low dimensional features have better visualization properties. PCA algorithm process: (Assuming the sample set has *m* samples, each with *n* features)

1) Form the original data into a matrix of $n \times m$.

2) Perform zero mean averaging on each row of the matrix, that is, subtract the mean of that row from each feature.

3) Calculate Covariance matrix

4) Find all eigenvectors and corresponding eigenvalues of the Covariance matrix.

5) Based on the eigenvectors corresponding to the largest to smallest eigenvalues, the first k eigenvectors are extracted to form a feature matrix u.

6) Rotate the original data into the space where the feature matrix u is located, $x = u^T x$, and the data obtained is the result of dimension reduction.

The PCA algorithm in this article reduces the dimensionality of previously strongly correlated variables to 10 principal components. And weighted sum these 10 principal components to obtain the final evaluation value, with the following weights representing the variance contribution rate of each principal component.

[9.99485148e-01, 2.03406676e-04, 1.54577892e-04, 1.49917693e-04, 3.32589571e-06, 1.96597261e-06, 1.35535425e-06, 2.18212458e-07, 6.98937483e-08, 1.19548550e-08]

3.4. Classification and discrimination based on SVM

Support vector machine is a binary classification model. Its basic model is the Linear classifier with the largest interval defined in the feature space. SVM also includes a kernel function, which makes it a virtually nonlinear classifier. Support vector machines have the characteristics of low generalization error rate, low computational overhead, and easy interpretation of results. This paper uses support vector machine to identify Insurance fraud in insurance claims database. The kernel function of support vector machine adopts Sigmaid Kernel, which can be used as a proxy for neural networks. The Activation function of the sigmoid core is a bipolar sigmoid function.

The sigmoid Gaussian kernel function formula is as follows:

$$k(x, y) = tanh(ax^T y + c) \tag{4}$$

Where, *tanh*(·) is the tangent function of Hyperbola, and a>0, c<0.

When different values are selected for the parameters of the SVM model, it will have different impacts on the classification results. This article analyzes the significant features of the insurance claims database, and the test results are shown in Figure 3. It can be seen that when the SVM parameter gamma=1, C=1000, the classification effect is the best.

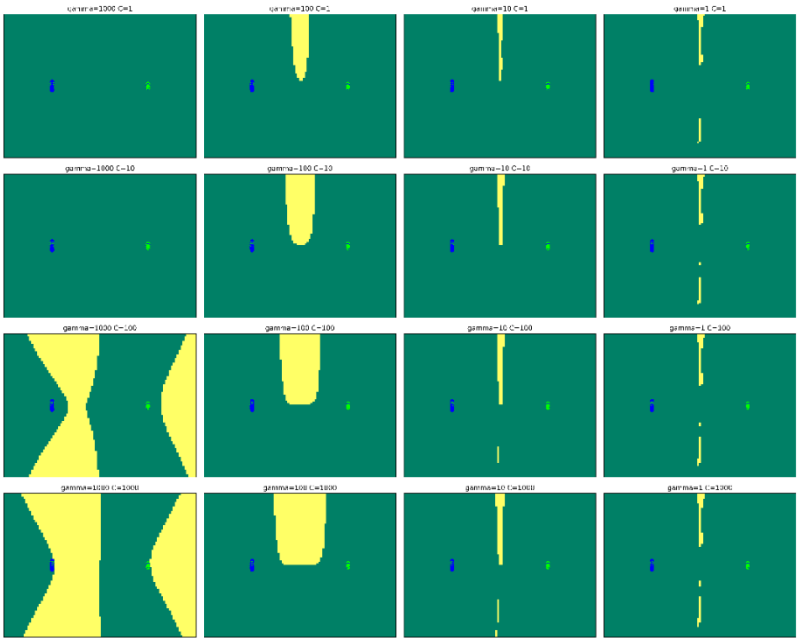

Figure 3. SVM Function Parameter Experiment

4. Results & discussion

This article focuses on 700 samples from the insurance claims database, dividing 490 samples into training set samples and 210 samples into test set samples. The sample set is divided into randomly selected samples.

The Discriminative model of Insurance fraud in this paper, after data preprocessing and extraction of strong correlation sample attributes of Insurance fraud, after SVM model training, the Receiver operating characteristic obtained on the validation dataset is shown in Figure 4.

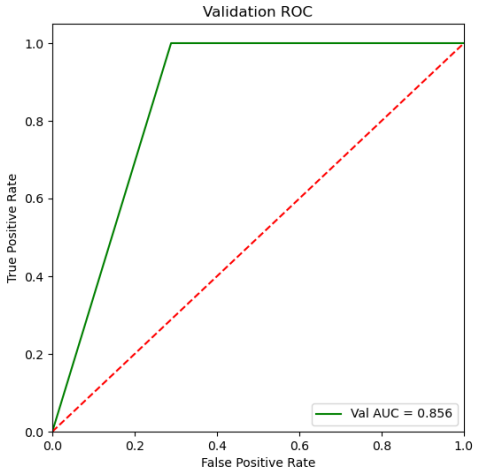

Figure 4. SVM insurance fraud discriminative model ROC curve

The classification report of SVM discriminative model for insurance fraud are shown in Table 1.

Table 1. Insurance Fraud SVM discriminative model Classification Report

	Precision	Recall	F1-score	Support
0	1.00	0.71	0.83	156
1	0.55	1.00	0.71	54
accuracy			0.79	210
macro avg	0.77	0.86	0.77	210
weighted avg	0.88	0.79	0.80	210

The AUC value of SVM classifier reaches 0.856. According to the existing classification standard of AUC value, the test of SVM model reaches a good level and can predict Insurance fraud behavior. The accuracy of this model is 0.79, and the recall rate is 0.71, which belongs to a good level. This shows that although SVM algorithm model has good Insurance fraud discrimination ability, there is still room for further improvement.

5. Conclusion

In a word, this paper studies the problem of Insurance fraud behavior discrimination modeling based on insurance claims database. The verification shows that Insurance fraud is related to the correlation of sample attribute variables. Through correlation analysis, only some attribute variables show strong correlation with Insurance fraud, while other attribute variables are interference terms of the Discriminative model. The focus of this article is to use principal component analysis algorithm to extract feature representations of strongly correlated attribute variables. Finally, support vector machine classifier is used to accurately judge Insurance fraud. In the future, the continuous improvement of Insurance fraud behavior discrimination model will help to avoid Insurance fraud, ensure the positive flow of insurance financial funds, and improve social stability.

References

[1] Q. Zhu. Feature Selection Based on the Discriminative Significance for Sparse Binary-Valued and Imbalanced Dataset [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2023, 37(03).

[2] A. Jeffrey, V. S. Caroline..Fraud detection in motor insurance: privacy and data protection concerns under EU Law. International Data Privacy Law, 2022(3): 3.

[3] R. Y. Gupta, S. S. Mudigonda, P. K. Baruah. TGANs with Machine Learning Models in Automobile Insurance Fraud Detection and Comparative Study with Other Data Imbalance Techniques. International Journal of Recent Technology and Engineering, 2021, 9(5): 236-244.

[4] M. Vogels, R. Zoeckler, D. M. Stasiw, et al. P. F. Verhulst's "notice sur la loi que la populations suit dans son accroissement" from correspondence mathematique et physique. Ghent, vol. X, 1838. Journal of Biological Physics, 1975, 3(4): 183-192.

[5] M. Artis, M. Ayuso, M. Guillen. Detection of Automobile Insurance Fraud with Discrete Choice Models and Misclassified Claims. The Journal of Risk and Insurance, 2002, 69(3): 325-340.

[6] D. E. Rumelhart, G. E. Hinton, R. J. Williams. PDP: Computational models of cognition and perception. Journal of Political Economy, 1986, 83-97.

[7] B. Botond, Ede Laszlo. Identifying Key Fraud Indicators in the Automobile Insurance Industry Using SQL Server Analysis Services. Studia Universitatis Babes-Bolyai Oeconomica, 2019, 64(2): 35-42.

[8] E. S. Pearson, B. A. S. Snow. Tests for rank correlation coefficients. Biometrika, 1962(1-2): 1-2.

[9] L. BREIMAN, J. H. FRIEDMAN, R. A. OLSHEN, et al. Classification and Regression Tree. Monterey, California, U.S.A.: Wadsworth International Group, 1984.

[10] C. Cortes, V. Vapnik. Support vector networks. Machine Learning, 1995, 20(3): 273-297.

[11] L. Breiman. Bagging Predictors. Machine Learning, 1996, 24(2): 123-140.