A Two-stage Fraud Detection Method

Yuanyuan Zhou, Wei Liu, Zhigang Hu and Feng He

***Abstract*—**With the economic development, more and more people participate in medical insurance and enjoy the benefits of medical insurance. However, medical insurance fraud has brought tremendous losses to the medical insurance fund. This paper presents a two-stage outlier detection method for medical insurance fraud detection. In the first stage, weighted k-means algorithm is used to cluster the data set and prune the result set, where the weight w is calculated by using the particle swarm optimization algorithm to minimize the evaluation function of the weight index. The second stage adopts the improved outlier detection method to process the result set. Experiments show that the accuracy of this method is higher than that of using K-means algorithm or LOF algorithm alone. Moreover, it can avoid the influence of subjective factors on the detection results.

***Keywords*—**medical insurance fraud, weighted k-means algorithm, particle swarm optimization, evaluation function of the weight index, outlier detection.

I. INTRODUCTION

Medical insurance fraud refers to the behavior of citizens, legal persons or other organizations that cause losses to the medical insurance fund when participating in medical insurance, paying medical insurance premiums, enjoying medical insurance benefits, deliberately fabricating facts, resorting to fraud, and concealing the real situation. Medical insurance fraud is causing huge losses to public health care funds around the world. Medicare fraud costs more than $ 8 billion in U.S. taxpayers annually, according to the U.S. Federal Bureau of Investigation (FBI) [1]. With the expanding medical insurance system in China, many illegal criminals extend black hand to the medical insurance fund. The means of illegal and criminal activities are becoming more and more concealed and the methods are constantly renovating. Even the insured person and the designated medical institutions collude to cheat and cheat together. Serious threat to the medical insurance fund. Fraud not only caused economic losses, but also seriously hindered the medical system in providing more quality and safe medical services to patients.

T. P. Hillerman [2] analyze abnormal medical claims by medical providers and obtained suspicious claims using CRIS-DM's research methodology. The literature describes in detail the choice of different eigenvalues and conducts experiments and analyzes. Reference [3] propose a scoring model to detect the fraud of electronic insurance claims in outpatient clinics. The model is divided into two stages, one is the degree of quantitative abuse of scoring, the other is dividing similar problem patterns. Two claims with high combined scores indicate a high likelihood of fraud. Reference [4] proposed a comprehensive fraud detection method, one is to analyze the user's behavior patterns to get the probability of fraud, the other is to use the improved LOF algorithm (SimLOF algorithm) to get the probability of fraud. The probability of fraud obtained by the two methods is combined with Dempster-Shafer Evidence theory, and the final result of fraud is obtained. The experimental data of the literature are extracted from the medical insurance system in Zibo City, Shandong Province. Experiments show that the method is 30% more efficient than the LOF method and the pattern recognition method. Q, Liu [5] selected the total amount of compensation and geographical location as the features, clustering analysis of three different disease data sets. The remote location and high consumption amount are regarded as fraudulent. Since there is no data that is actually labeled as fraudulent, the paper simply finds the data of high suspiciousness theoretically and analyzes the rationality of the result. In this paper, only two features are chosen and the result is analyzed simply by selecting the least number of clusters as fraud. So the accuracy is very low. The choice of the number of clusters which completely determined by the artificial, it's too subjective.|

This paper proposes a two-stage fraud detection method that can detect specific type of fraud. In the first stage, a weighted K-means clustering method is used to cluster data with the same characteristics and set a threshold to reduce the clustering result set. Among them, the weight value is calculated by particle swarm optimization algorithm. This results in a set of eigenvalues that depend only on the given data. To avoid the subjectivity of K value selection, the value of K is selected by 26 comprehensive indexes. In the second stage, an improved outlier detection method is used to obtain fraudulent data according to the outlier score. The experiments results are compared with the K-means clustering algorithm and LOF algorithm, and the high accuracy of the method is proved.

This work was supported by Central south university graduate student independent exploration innovation project funds of China under grant No.2017zzts603, the national natural science funds of China under grant No.61572525.

Yuanyuan Zhou is with Software College, Central South University, Changsha 410075, Hunan, China.

Zhigang Hu is with the Software College, Central South University, Changsha 410075, Hunan, China (corresponding author; e-mail: zghu@csu.edu.cn).

Wei Liu is with the School of Management and Information Engineering, Hunan University of Chinese Medicine, Changsha 410208, Hunan, China.

Feng He is with Software College, Central South University, Changsha 410075, Hunan, China.

II. Methodology

A. Method Process

The fraud detection method proposed in this paper is based on the methodology of data mining. In recent years, data mining has been used more and more in medical insurance fraud detection [6]. Firstly, the data is preprocessed and the features are selected according to the specific fraud types. Particle swarm optimization algorithm is used to find the optimal weight, and then weighted k-means is used to cluster. Outlier detection algorithm is used to score the results. Finally, the experimental results are analyzed. The process shown in Fig. 1:

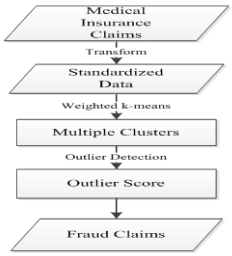

Fig. 1. Method flow diagram

B. Data Preprocessing

The critical step to do is data preprocessing. The structure and content of Medicare data itself are complicated. There is a lot of redundant information and noise, which seriously affects the efficiency of data analysis. How to reduce the dimension of medical insurance data and choose the feature is also an important issue in Medicare data analysis.

Data preprocessing includes data cleaning, data integration, data selection, data transformation. These are different forms of data preprocessing. Data cleaning is to eliminate noise and inconsistent data, data integration is to combine multiple data sources together, data selection is to extract data from the data related to the analysis task, data transformation is to transform or unify the data into a suitable mining form.

The above steps do not necessarily have to be used, one of the most important step is the data selection. According to the mining task, we choose the characteristic attribute set to facilitate the data mining. There are several commonly used methods for selecting feature attribute sets, including data cube aggregation, wavelet transform, attribute subset selection, principal component analysis and so on. Principal Component Analysis (PCA) is computationally expensive and can handle sparsely sloped data. It can treat multidimensional data as K-dimensional problems. Principal components can also be used as inputs to cluster analysis. The process of data preprocessing used in this paper is shown in Fig. 2:

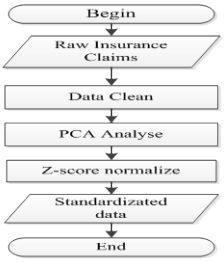

Fig. 2. Data preprocessing flowchart

III. Cluster Stage

K-means clustering is a commonly used method, which was refined by J. Hartigan and A. Wong in 1975 [7]. K-means clustering algorithms are relatively scalable and efficient in handling large data sets.

The algorithm stops when the criterion function converges. The criterion function usually adopts the square error criterion, which is defined as follows:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i| \tag{1}$$

The algorithm describes the degree of similarity of the sample point data Xp and Xq on the data set X. The commonly used Euclidean distance is defined as (2):

$$d_{pq} = \sqrt{\sum_{k=1}^{m} (x_{pk} - x_{qk})^2} \tag{2}$$

Which x_{pk} is x_p the kth dimension attribute. From this definition, we can clearly see that the algorithm considers all the attributes as equal and does not consider the influence of different attributes in practical applications on a specific problem. Therefore, improving the Euclidean distance is called weighted Euclidean distance and is defined as (3):

$$d_{pq}^{(w)} = \sqrt{\sum_{k=1}^{m} w_k^2 (x_{pk} - x_{qk})^2} \tag{3}$$

Where w_k is the weight value of the kth dimension attribute.

A. Attribute Weight Evaluation Function

In this paper, the Euclidean distance-based similarity independent method and the attribute weight evaluation function (abbreviated as $cf(w)$) are given in [8]. The specific definition is as (4) and (5):

$$\rho_{pq}^{(w)} = \frac{1}{1 + \beta d_{pq}^{(w)}} \tag{4}$$

$$cf(w) = \frac{-1}{n(n-1)} \sum_{p<q} [\rho_{pq}^{(w)} \log \rho_{pq} + (1 - \rho_{pq}^{(w)}) \log(1 - \rho_{pq})] \tag{5}$$

Reference [8] proposed that objects satisfying similar ($\rho_{pq}^{(w)} >$ 0.5) are more similar ($\rho_{pq}^{(w)} \to 1$) and dissimilar ($\rho_{pq}^{(w)} <$0.5) objects are more dissimilar ($\rho_{pq}^{(w)} \to 0$) when this function takes the minimum value.

B. Particle Swarm Optimization

There are many methods to solve the minimum value of

$cf(w)$. This paper uses particle swarm optimization algorithm to solve it. Particle swarm optimization algorithm starting from the random, iterative search for the optimal solution. It is simpler than genetic algorithm rules and easy to implement, with high accuracy and fast convergence. Reference [9] introduced particle swarm optimization in detail. Briefly introduce the particle swarm optimization algorithm is as follows:

Assuming a D-dimensional target search space, N particles form a community, where the ith particle is represented by a D-dimensional vector xi = $\{x_{i1}, x_{i2}, .. x_{iD}\}$, i = 1,2...N; Its "flying" speed is also a D-dimensional vector denoted as vi = $\{v_{i1}, v_{i2}, ... v_{iD}\}$, i = 1,2, ... N; The optimal position of the first particle so far is called individual extremum, which is denoted by pbest = $\{p_{i1}, p_{i2}, .. p_{iD}\}$, i = 1, ... N. The optimal position searched by the entire particle swarm so far is called global extremum, gbest = $\{p_{g1}, p_{g2}, .. p_{gD}\}$, g = 1, ... N; Particles keep track of pbest and gbest in flight and update their speed and position according to (6) and (7):

$$v_{id} = w v_{id} + c_1 r_1 (p_{id} - x_{id}) + c_2 r_2 (p_{gd} - x_{id}) \quad (6)$$

$$x_{id} = x_{id} + v_{id} \quad (7)$$

The w indicates the weight of the previous moment, r_1 and r_2 is a random number between 0 and 1. c_2 and c_1 is a learning factor.

C. Solution w

The algorithm steps of using particle swarm optimization and attribute weight evaluation function to solve w are as follows:

(1) Initializing population size N, particle velocity vi, particle position xi.

(2) The value of all particles calculated $cf(w)$ as their own fitness values as $f_{it}(i)$.

(3) Compare the fitness values of each particle $f_{it}(i)$ with the individual extremun $p_{best}(i)$, if $f_{it}(i) < p_{best}(i)$, then update the value of $p_{best}(i)$ as $f_{it}(i)$, that is $p_{best}(i) = f_{it}(i)$.

(4) Compare the fitness values of each particle $f_{it}(i)$ and the size of the global extremum $g_{best}(i)$, if $f_{it}(i) < g_{best}(i)$, then set $g_{best}(i) = f_{it}(i)$.

(5) Each particle updates its vi and xi according to the formula (6) and (7).

(6) Meet the minimun error exit the loop, otherwise the second step will be returned.

IV. Outlier Detection Stage

Since medical insurance fraud is only the behavior of a few reimbursors, we can a threshold M can be set to prune the result of the initial cluster. Clusters whose numbers are greater than M are directly excluded, leaving clusters with fewer than M clusters. Outliers are detected for these clusters, and fraudulent data is found.

Outlier mining has a wide range of applications that can be used for fraud detection by detecting unusual credit card usage or telecommunications services. There is a clear difference between the medical fraud behavior and the normal reimbursement medical behavior. This is the outlier, so the outlier detection algorithm is used.

Outlier detection algorithms include outlier detection based on statistical distribution, outlier detection based on distance, local outlier detection based on density, and outlier detection based on deviation. Outlier detection based on statistical distribution does not ensure that all outliers are detected. Outlier detection based on distance requires that users set up two important parameters and need to try out the trial many times [10]. Density-based local outlier detection can detect global outliers and local outliers simultaneously. The size of the outlier indicates the outlier. In this paper, the average k-distance as a measure of the outlier, is now defined as (8):

$$f(x_i) = \frac{\sum_{q \in N_k(x_i)} dist(x_i, q)}{\left| N_k(x_i) \right|} \quad (8)$$

The larger the value $f(x_i)$, the greater the extent of the outlier. Where $dist(x_i, q)$, denotes the distance of object q and xi, q is in the neighborhood of k distance of xi. Here also introduce the k distance and k distance neighborhood.

The k distance of object p, is the maximum distance from p to its k nearest neighbor, denoted by $k-dis\tan ce(p)$. Two conditions need to be fulfilled:

(1) There are at least k objects o' ∈ D, with dist(p, o ') ≤ dist(p, o).

(2) There are at most k-1 objects o " ∈ D, with dist(p, o") < dist(p, o).

K distance neighborhood is denotes as $N_{k-dis\tan ce(p)}(p)$ and is abbreviated as $N_k(p)$, it represents all objects at a k-distance no greater than p.

V. Experiment

A. Data

The experimental data in this paper is a one-month reimbursement data collected by Shenzhen Medicare Center with a total of 300,000 pieces of data. According to our analysis, fraud within a month may be expressed in the form of high bills and more times of taking medicine.

In order to achieve the purpose of this paper for the detection of specific types of fraud, we cleaned and converted the raw data. We chose a feature sets to fit PCA analysis. After the analysis, some important features are left. The final features are: billing number, patient id, age, times of take medicine, office, fee.

Data were normalized to [-1, 1] using z-score method. After preprocessing, there were 50,000 data used for cluster analysis, of which 116 were manually verified as fraudulent, about 0.2% of the total. Weights obtained using the PSO algorithm, department weight: 2.96, times of take medicine weight: 2.61, total weight: 2.91, age weight: 2.41.

B. Determine the K value

The choice of K value is usually defined manually. To avoid

the influence of subjective choice on the experimental results, the optimal K value is determined by the NbClust function [11], as shown in Fig. 3, which indicating that 46 clusters are a suitable number.

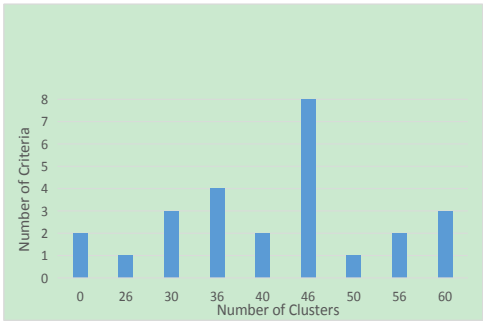

Fig. 3. Number of Clusters chosen by 26 Criteria

C. Clusters

After clustering, each data is grouped into the corresponding cluster according to their similarity. Fig. 4 shows the number of clusters.

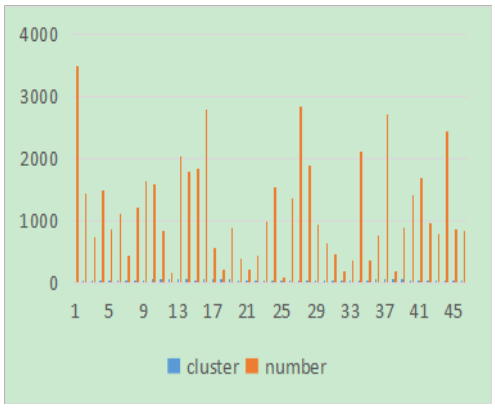

Fig. 4. Number of clusters and clusters

Set the threshold M = 1000. The clusters whose number is greater than 1000 don't participate in outlier detection stage. Because they are legal reimbursement data. There are 26 clusters used for outlier analysis after culling.

D. Outlier score

The data in these 26 clusters are processed with an improved outlier detection algorithm. The top 10 outlier scores are shown in Table 1:

Table. 1. From the group of top ten records

De-indentifed Bill ID	Outlier-ness Score
1	9.6783
2	9.6763
3	9.6583
4	8.9653
5	8.7553
6	8.6453
7	8.6353
8	8.5723
9	8.5563
10	8.5453

E. Analysis

Outliers score above 5 points are considered to be suspected of fraud. A total of 84 records were selected. Return the results to the initial data, some of the data shown in Table 2.

As can be seen from the table, the times of taking medicine with serial numbers 1, 2, 3 and 9 is less, but the total cost of individual bills is very high. The patients with serial numbers 4, 5, 6, 7, 8 and 10 received more drugs, and the cost of a single bill is also high. These experimental results are consistent with the target of fraud detection. It can be seen from the experimental results that the method is more efficient in detecting the audited reimbursement data.

Table. 2. Part of the record after clustering

ID	BillNum	PatieID	Medic-times	Office	Fee	Age
1	5354736	689520	1	329	1125.26	38
2	5253872	217527	2	129	1884.3	40
3	5258160	267254	2	112	1480.0	42
4	5101180	656502	4	329	1315.21	31
5	5126120	178614	9	203	1226.99	30
6	5273220	178614	9	203	1126.98	42
7	5101764	656502	4	329	1129.36	51
8	5380464	173602	10	203	1104.45	48
9	5267933	581961	1	10	1286.49	47
10	537156	591234	15	10	2318.91	46

73 of the 84 records were previously audited as fraudulent records. The traditional K-means algorithm and LOF algorithm are respectively used to conduct experiments on this datasets, and their results are compared with the method in this paper. Algorithm accuracy as shown in Table 3:

Table. 3. Fraud identification accuracy

Method	Accuracy(%)
Two-stage method	87
K-means	78
LOF	75

VI. Conclusion

This paper presents a two-stage detection method to detect unusual claim behaviours. First introduced attribute evaluation function, using particle swarm optimization algorithm to get the best weights. Then using the weighted k-means algorithm to cluster grouping, set a threshold value M for pruning the clustering results. Again with the improved the outlier detection algorithm based on density of outlier analysis, get the final result.

In order to verify the effectiveness of the method, using real-world medical insurance reimbursement data to do the experiment. The dataset contains more than 30 million medical claim activity records. The experiment results proved that the proposed approach could achieve higher accuracy than existing fraud detection methods.

The subsequent research, we will be able to consider the most common cases of medical insurance fraud crime, excavate the medical reimbursement staff visit mode. A detection scheme will be proposed in order to deal with the changing means of fraud.

References

[1] N. Aldrich, J. Crowder, and B. Benson, "How much does medicare lose due to fraud and improper payments each year", *The Sentinel*, 2014.

[2] T. P. Hillerman, RN. Carvalho, and ACB. Reis, "Analyzing Suspicious Medical Visit Claims from Individual Healthcare Service Provider Using K-means Clustering", *Electronic Govrnment and the Information System Perspective, Springer International Publishing*, 2015, pp. 191-205.
[3] H. Shin, H. Park, and J. Lee, "A scoring model to detect abusive billing patterns in health insurance claims", *Expert System with Applications*, vol. 39, no. 8, pp. 7441-7450, 2012.
[4] C. F. Sun, Q. Z. Li, and L. Z. Cui et al, "An Effective Hybrid Fraud Detection Method", *International Conference on Knowledge Science, Engineering and Management*, Springer-Verlag New York, Inc, 2015, pp. 563-574.
[5] Q. Liu, M. Vasarhelyi, "Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information", *in Proc. 29th. World Continuous Auditing and Reporting Symposium*, Brisbane, 2013.
[6] H. Joudaki, A. Rashidian, B. Minaei-Bidgoli, "Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature", *Global Journal of Health Science*, vol. 7, no.1, pp. 194, 2014.
[7] J. Hartiga, A. Wong, "Clustering algorithms", Wiley, New York, 1975.
[8] L. Wang, S. Guan, X. Wang, "Fuzzy C Mean Algorithm Based on Feature Weights", *Chinese Journal of Computers*, vol. 29, no. 10, pp. 1797-1803, 2006.
[9] Y. Zhang, S. Wang, G. Ji, "A Comprehensive Survey on Particle Swarm Optimization Algorithm and Its Applications", *Mathematical Problems in Engineering*, vol. 2015, no. 1, pp.1-38, 2015.
[10] G. V. Capelleveen, M. Poel, J. V. Hillegersberg, and R. M. Mueller, "Outlier-based Health Insurance Fraud Detection for U.S. Medicaid Data", *International Conference on Enterprise Information Systems*, SCITEPRESS - Science and Technology Publications, Lda, 2014, pp. 684-694.
[11] M. Charrad, N. Ghazzali, V. Boiteau, "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set", *Bmc Health Services Research*, vol. 61, no. 6, pp. 1-36, 2014.