Discovery of fraud in Medical Insurance

Yajie Cai[1], Zhe Yin[2*]

Department of Mathematics,Yanbian University,CHINA

Abstract— *The term insurance fraud refers to the commission of any act with the intent to obtain an outcome that is favorable, but fraudulent during an insurance claim. Including single prescription medicines is extremely high, card repeatedly within a certain amount of time for medicine, etc. This paper is based on methods of hierarchical cluster analysis and generalized squared distance discriminate method to record medical coverage of transaction data at outliers for finding out the corresponding abnormal record which indicates potential fraud.*

Keywords— medical insurance fraud, hierarchical cluster analysis, training sample, generalized square distance discrimination.

I. INTRODUCTION

Insurance fraud is very serious and widespread violations in the insurance industry. This paper studies the methods of data preprocessing, visualization, hierarchical cluster analysis and generalized squared distance discrimination.

II. QUESTIONS AND STEPS

First of all, processing to the patient ID for the index contains complete information summary, the patient with graphics and descriptive statistics to said patient information for each attribute, to detect the general information of data, describe its profile information. Secondly, on the basis of applying the theory of cluster analysis to generate training samples. Finally, classify the original objects according to the criterion which relay on generate training samples, so we can successfully extract the abnormal points, namely the potential fraud of patient information.

For the first problem, we can infer the age of the patient according to the information given in the annex (In January 31, 2016 as the end of calculation) and make summary information table. For the convenience of processing information, frequency of patient ID which appears at different time is used as a measure of index. Through the quantitative analysis of relevant properties, complete description of the gender, age, frequency, and the inherent law of the information such as the total price.

For the second problem, based on the information obtained from the first problem, considering the modeling process of simplicity and intuitive, clustering by using system theory[2] and statistical software SAS, the original sample by using simple random sampling to sample clustering (Q type cluster), generates a capacity of 10000, the training samples of 4 categories.

For the third problem, based on the second clustering of the training sample, using the SAS to process data, as a criterion for judging the discriminate classification to the original object, and thus the abnormal samples are extracted successfully points, namely, what we're looking for potential fraud of patient information.

Step:

In order to find out the existing medical insurance fraud [1], we take the following steps:

The first step: Use Excel (vlookup function and data perspective function) to collect a complete patient information table, and analyze the patient's information for each attribute description.

The second step: According to the conclusion from problem one, using the SAS to sampling from the original sample. We classify the samples based on the theory of system clustering. And determine the number of classes according to the hierarchical graph and statistics generating a certain capacity of convenient follow-up studies of training sample.

The third step: According to the question 2 to get the training sample, determine the overall number of classes. And using the training samples as the criterion for judging the discriminate classification of the original object data, find out the abnormal point in the sample, which is potentially fraudulent behavior of patient information.

III. SOLVE THE PROBLEM ONE

Based on the known information, the hidden information of the patient is inferred, and the detailed table of all the information of the patient is obtained.

3.1 Age and Gender

According to the known data infer a patient's age, for the convenience of information processing and using different time id appears the frequencies for the purchase frequency is a measure, making the frequency of age and gender, percentage, cumulative frequency and cumulative percentage of the map, as shown in Figure 3-1:

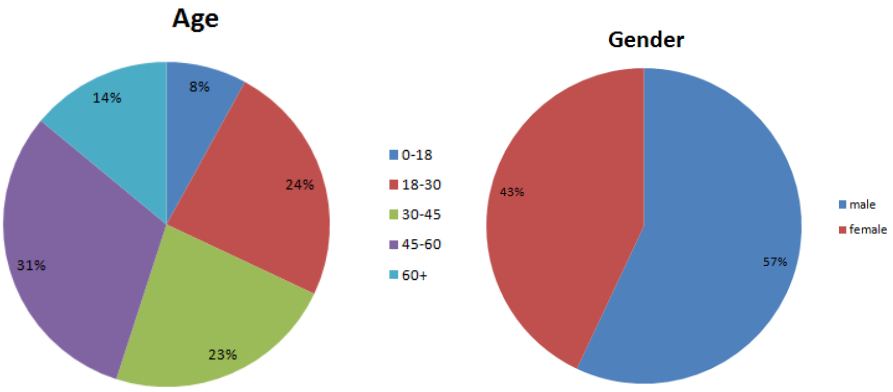

FIGURE 3-1 AGE, GENDER, AND SCALE MAP

TABLE 3-1
THE FREQUENCY AND PERCENTAGE TREATMENT TABLE

Age	Frequency	Percentage	Cumulative frequency	Cumulative percentage
0-18	13894	23.95	13894	23.95
18-30	13413	23.12	27307	47.07
30-45	18226	31.42	45533	78.49
45-60	7966	13.73	53499	92.22
60-∞	4513	7.78	58012	100.00

TABLE 3-2
THE FREQUENCY AND PERCENTAGE TREATMENT TABLE

Gender	Frequency	Percentage	Cumulative frequency	Cumulative percentage
Female	32918	56.74	32918	56.74
Male	25096	43.26	58014	100.00

Through the observation of tables 3-1and 3-2, we could draw the conclusion. The ages of 30 to 45 patients has a greater frequency, there is nearly no difference between the ages of 0 to 18 and 18 to 30. Frequency of minimum age distribution area of 60 and above, it's in our objective conditions. Given the gender differences, we found that the proportion of women more than men and women with greater frequency.

3.2 Prices

Consider the differences value of the medicine for the purchase, we use Excel to draw the picture for analyzing the distribution of drug prices and the frequency distribution of different prices. As shown in figure 3-2:

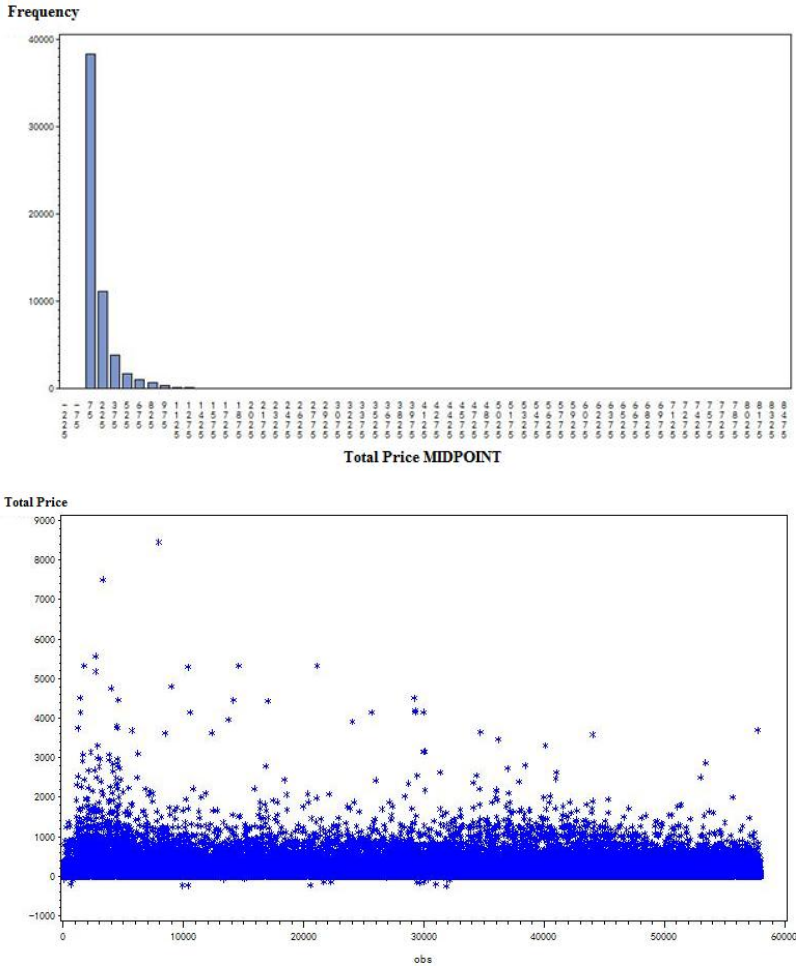

FIGURE 3-2 PURCHASE PRICE DISTRIBUTION OF THE MEDICINE

According to the chart above, we could know that the drug which was bought by patient has the largest sum frequency between the 75 to 225. Among them, the price between the 75 to 975 can be regarded as normal price. Distribution of the price of more than 975 patients has less, lower frequency, deviating from the overall cost level, so there may be a health insurance fraud.

We can get a contour map of different drug price (Total price), frequency, drug purchase quantity (Number), as shown in figure 3-3 (small black dots distribution). It's not difficult to find patients mainly distributed in the region of the low price, low frequency and low quantity (that is, the figure in the lower left corner area), while other regional distribution is relatively sparse, especially high price and high quantity, low frequency area. We will take it as a key monitoring object.

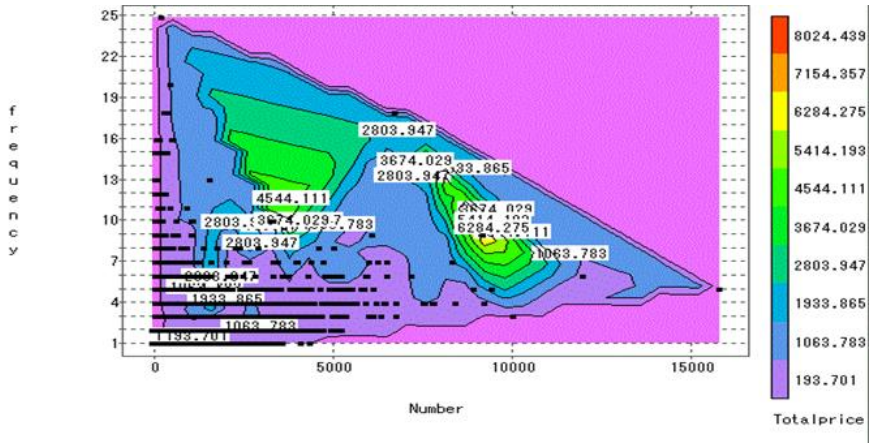

FIGURE 3-3 PATIENTS' SITUATION MAP

IV. SOLVE THE PROBLEM TWO

There're a total of 58014 patients overall information, we decided to use simple random sampling (by proc surveyselct process step implementation) from a sample of 10000 patients information in general observations. From four indexes of each sample (gender, age, and The Times of price), we can get the observed data $x_{ij}(i=1,...10000;\ \ j=1,2,3,4),$ which is shown in the table below.

TABLE 4-1
MATRIX OF THE PATIENT DATA

	X_1	X_2	X_3	X_4
$X_{(1)}$	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$
$X_{(2)}$	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$
…	…	…	…	…
$X_{(10000)}$	$x_{10000,1}$	$x_{10000,2}$	$x_{10000,3}$	$x_{10000,4}$

The data matrix of the patient data in the table is represented as:

$$X=\begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} \\ x_{2,1} & x_{2,2} & x_{2,3} & x_{2,4} \\ \vdots & \vdots & \vdots & \vdots \\ x_{10000,1} & x_{10000,2} & x_{10000,3} & x_{10000,4} \end{bmatrix} \tag{1}$$

In the formula, the column vector $X_j=\left(x_{1,j},x_{2,j},...,x_{10000,j}\right)$ represents the index of $j\ (j=1,2,3,4)$ while the row vector $X_{(i)}=\left(x_{i,1},x_{i,2},..x_{i,3},x_{i,4}\right)$ represents the information of the patient i .

The basic steps of system clustering [3] are given as follow:

(1) Data conversion

In order to facilitate comparison and to eliminate the influence of the dimension, we first need to carry on the standardized transformation of the data before making the cluster:

$$x_{ij}^{*}=\frac{x_{ij}-\overline{x}_j}{s_j}\ \ i=1,2,...10000;j=1,2,3,4 \tag{2}$$

The $\overline{x}_j$ is given by:

$$\overline{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij};\ s_j^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij}-\overline{x}_j\right)^2;\ j=1,2,3,4$$

After the transform, the sample mean of each variable is 0, the standard deviation is 1, and the normalized transformed data $\left\{x^{*}_{ij}\right\}$ is independent of the dimension of the variables.

(2) Calculate the distance between the two samples

Select the Euclidean distance :

$$d_{ij} = \sqrt{\sum_{t=1}^{m} \left| x_{it} - x_{jt} \right|^2} \ (i, j = 1, 2, .., n) \tag{3}$$

(3) Clustering process

Firstly, we know that the n samples respectively comprise one, so the number of classes $k = n:\ G_i = \left\{ X_{(i)} \right\}\ (i = 1, ..., n)$ the distance between classes is the distance between samples ($D^{(1)} = D^{(0)}$).

Let $j = 2, ..., n,$ to perform the following class processes

- Merging the minimum cluster distance between two classes for a new class (the minimum cluster distance between classes by using McQuitty similarity analysis method [4])

The total number of classes is reduced by one class at this time, that is $k = n - j + 1$

Set a step, letting G_K and G_L merge to G_M, for any kind of G_J, consider the triangle which the side length is consist of D_{KL}, D_{LJ} and D_{KJ} (Figure 4-1), take the edge of D_{KL} centerline as D_{MJ}. By the elementary plane geometry, the formula for D_{MJ} is given as:

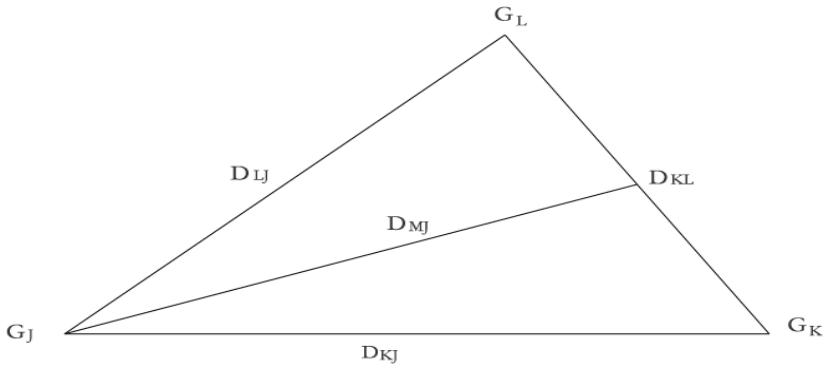

FIGURE 4-1

$$D_{MJ}^2 = \frac{1}{2} D_{KJ}^2 + \frac{1}{2} D_{LJ}^2 - \frac{1}{4} D_{KL}^2 \tag{4}$$

In the formula (4), let the coefficients changed to the parameters with the parameters β , just as follows:

$$D_{MJ}^2 = \frac{1\text{-}\beta}{2} (D_{KJ}^2 + D_{LJ}^2) + \beta D_{KL}^2 \tag{5}$$

When β <1, we called this method as flexible method. When β =0, the recursive formula is changed into

$$D_{MJ}^2 = \frac{1}{2} (D_{KJ}^2 + D_{LJ}^2) \tag{6}$$

- Calculate the distance between new class and other classes to get a new distance matrix $D^{(j)}$

If the total number of the merged class is still more than 1, then repeat step 1) and 2) until the total number of the class become 1

(4) Spectrum diagram as shown in figure 4-2

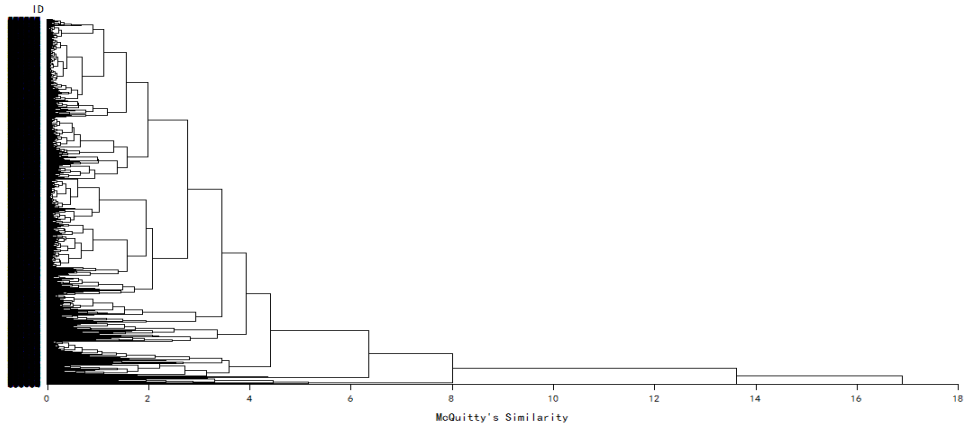

FIGURE 4-2 SPECTRUM DIAGRAM

Because the observation is too much so the tree figure appears cluttered. From the figure we can also see the clustering situation for each sample point at different levels. When the sample is divided into 4 classes, it can share very open.

(5) Determine the number of categories according to statistics [3]

In the CLUSTER process of SAS, the method of determining the number of classes by statistics is derived from the statistical analysis of variance:

- pseudo F-statistics：

$$pseudo\ \mathrm{F}-statistics=\frac{(T-P_G)/(G-1)}{P_G/(n-G)} \tag{7}$$

The evaluation of pseudo F-statistics is divided into G class effects. The larger the pseudo F-statistics, the more reasonable the .classification Often take the clustering level of large amount of pseudo F-statistics and smaller class number.

- $Pseudo\ T^2-\ statistic$ ：

$$Pseudo\ T^2-\ statistic=B_{KL}/((W_K+W_L)/(n_k+n_L-2)) \tag{8}$$

This statistic is used to evaluate the effect of the combined class C_K and the class C_L. The value of this value indicates that the two classes are very separate and should not be combined with these two classes.

- R^2 statistics：

$$R^2=1-\frac{p_G}{T} \tag{9}$$

The larger the R^2, the smaller the sum of squared residuals in each class when we divided into G class, which means such classification is reasonable. However, with the increase of the classification, each class will get smaller while R^2 will get greater, so we can only take G class to make R^2 is large enough, but the class itself is small and the R^2 is no longer greatly increased.

●Semi-partial R^2 statistics:

When merging class C_K and class C_L to class C_M, definite semi-partial R^2 statistics:

$$\text{Partial statistic } R^2 = \frac{B_{KL}}{T} \tag{10}$$

$B_{KL} = W_M - (W_K + W_L)$ is the increment of sum of deviation square within the class, which caused by merging class. Use semi-partial R^2 for evaluation of the effect of a merger, the value is the difference between the last R^2 step and the step R^2.The greater the value, the better the effect of the last merger.

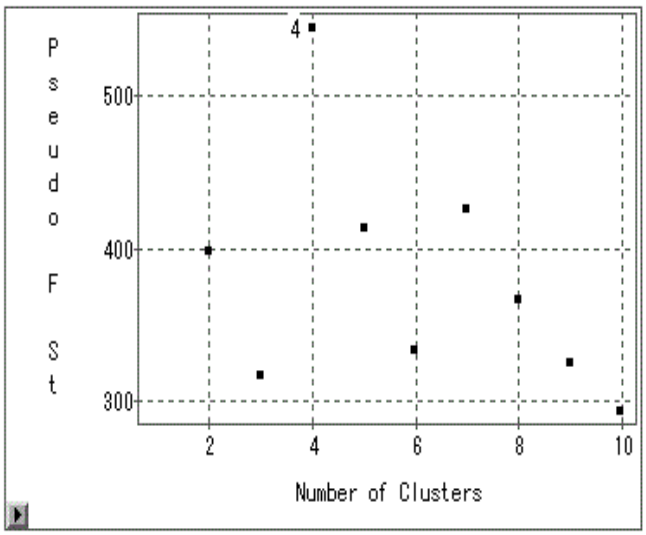

FIGURE 4-3 PSEUDO F STATISTICS

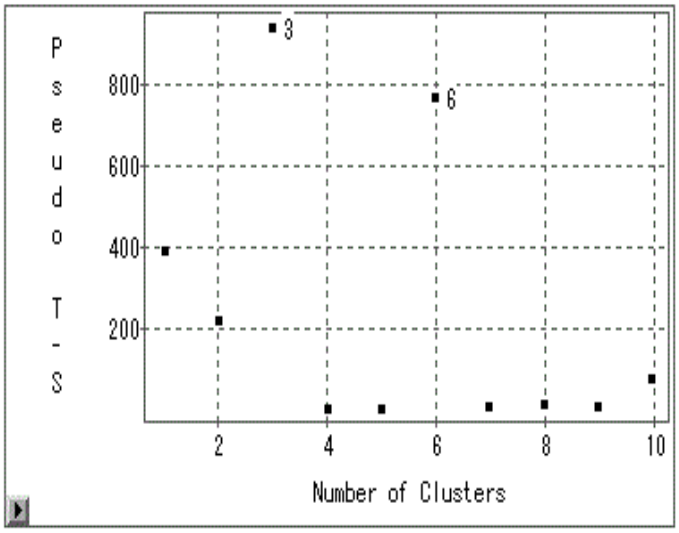

FIGURE 4-4 PSEUDO t^2 STATISTICS

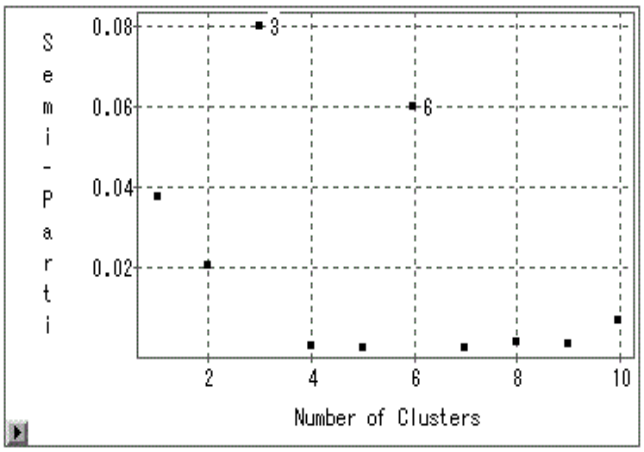

FIGURE 4-5 R^2 STATISTICS

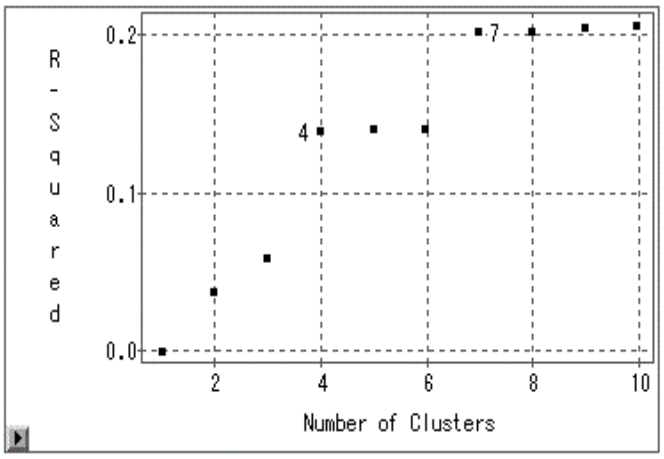

FIGURE 4-6 SEMI-PARTIAL R^2 STATISTICS

The statistics of the graph is shown in Figure 4-3 to figure 4-6:

● The maximum value of the pseudo F statistic is NCL=4, which indicates that it is more appropriate to divide into 4 kinds according to pseudo F criterion.

- The maximum and second large values of pseudo t^2 statistics are NCL =3 and 6 (local), that is, the effect of the last clustering is better. According to the pseudo t^2 criteria, it is appropriate to divided into 4 or 7 categories.

- In the 3 categories and 4 categories, as well as the 6 and 7 categories, the process of R^2 decreased more. By analyzing the R^2 statistics, it can be concluded that it is more suitable to be divided into 4 and 7 categories.

- The largest and second largest semi-partial R^2 statistics were NCL =3 and 6, which means it is appropriate to divided into 4 and 7 categories according to the semi-partial criteria.

It can be seen that the pseudo F statistic support to divide into 4 categories while the standards of CCC、pseudo t^2、pseudo R^2 and semi-partial R^2 support the category 4 and 7. The comprehensive analysis shows that the 10000 samples are divided into 4 or 7 categories by McQuitty similarity analysis method. Select the 4 category for the final number of samples, the classification of the sample data set as a discriminate analysis of the training samples. The classification results are as follows:

TABLE 4-2
FREQUENCY AND PERCENTAGE OF EACH CATEGORY

Cluster	Frequency	Proportion	Prior
1	9822	0.982200	0.25
2	169	0.016900	0.25
3	3	0.003000	0.25
4	6	0.006000	0.25

V. SOLVE THE PROBLEM THREE

The above process through cluster analysis initially established a training sample, through training samples to establish a criterion. For the whole of the 58014 samples, based on this criterion to distinguish which classification of samples it comes from. The distance of samples X to the generalized squared G_t is given as:

$$D_t^2(X) = D^2(X, G_t) = d_t^2(X) + g_1(t) + g_2(t), \tag{11}$$

Let,

$$g_1(t) = \begin{cases} \ln|S_t|, & \exists_{i,j}\ s.t \quad \partial_i \neq \partial_j \\ 0, & \forall_{i,j}\ s.t \quad \partial_i = \partial_j \end{cases} \tag{12}$$

$$g_2(t) = \begin{cases} -2\ln|q_t|, & \text{If a priori probabilities are not all equal} \\ 0, & \text{If a priori probabilities are all equal} \end{cases} \tag{13}$$

S_t is the group covariance matrix of class t. When the time $D_t^2(X) < D_i^2(X)$, $(i \neq t, i = 1,...,k)$, let $X \in G_t$. Because the prior probabilities are known, in the following criteria we specify priors equal.

The DISCRIM process of SAS software is used to determine the generalized square distance, and the output results are as follows:

TABLE 5-1
GENERALIZED SQUARED DISTANCE BETWEEN VARIOUS TYPES

Generalized Squared Distance to type				
From type	**1**	**2**	**3**	**4**
1	0	30.14314	486.93062	299.48602
2	30.14314	0	278.49218	298.41022
3	486.93062	278.49218	0	750.54533
4	299.48602	298.41022	750.54533	0

TABLE 5-2
DISCRIMINANT CLASSIFICATION RESULTS OF TRAINING SAMPLES

Number of Observations and Percent Classified into CLUSTER					
From cluster	**1**	**2**	**3**	**4**	**Total**
4	0	0	0	6	6
	0.00	0.00	0.00	100.00	100.00
Total	9622	354	4	20	10000
	96.22	3.54	0.04	0.20	100.00

TABLE 5-3
ERROR RATE FOR EACH CATEGORY

Error Count Estimates for CLUSTER					
	1	2	3	4	Total
Rate	0.0204	0.0059	0.0000	0.0000	0.0066
Priors	0.2500	0.2500	0.2500	0.2500	

Table 5-1 shows the various types of generalized squared distance, it's not difficult to find that the distance between the classes square is less than the class distance;

Table 5-2 shows the classification summary result of training samples back to the generation. From each category (i.e. the cluster analysis generated classes) of sample by discriminant function included in all kinds of frequency and percentage (a frequency and a percentage).

Table 5-3 shows the error count estimates for cluster in each category, it can be seen that the class 1 has the highest fault point rate, which is 0.0204. Each category are always in the wrong points rate estimates for 0.0066 (actual total misclassification rate is greater than the value).

TABLE 5-4
CLASSIFICATION RESULTS OF TOTAL SAMPLES

Number of Observation and Percent Classified into CLUSTER					
	1	**2**	**3**	**4**	**Total**
Total	55781	2104	30	97	58012
	96.15	3.63	0.05	0.17	100.00
Priors	0.25	0.25	0.25	0.25	

The results of overall classification are in the following table. The class 3 and 4 for outliers in the normal range of deviation from the overall, which is potentially fraudulent behavior of patient information data set, so the class 3 is the key suspect (a total of 97 patients).

By the original data, it is shown that the Medicare Handbook No. 1 corresponds to different patient ID; this does not accord with the actual behavior. Here, we will be classified for did not participate in the Medicare population, and the list of patients excluded from doubt. Finally, we get the information of 75 patients with potential fraud patients), slightly.

VI. CONCLUSION

In this paper, the samples are divided into one class by cluster analysis, after that, the most similar two samples are clustered into a small class, and then merge with the most similar small classes. The Euclidean distance we used is not affected by the dimension. The Euclidean distance between two points has nothing to do with the measuring unit of the original data. The distance is the same between the two points which is calculated by the standardized data and the center data. This discovery has a certain reference value for the discovery and research of medical insurance fraud.

REFERENCES

[1] Xia hong,wang kai,zhang shouchun.Medical Insurance Fraud[J], 2007.

[2] Fan ming,fan hongjian et Interpret,Introduction to Data Mining.beijing:People Post Press,355-497.

[3] Wang yuanzheng,xu yajing.SAS Software and statistical applications Course[M]. Beijing:Mechanical Engineering Press, 2007.1.

[4] Xue yi,chen liping. Statistical modeling and software R[M]. beijing：Tsinghua University Press,2007.4,403-420.