Building of a Standardized Health Insurance Monitoring Model Based on Data Mining

Zhou Ji[1]

1 Second Affiliated Hospital School of Medicine
Zhejiang University
Hangzhou, Zhejiang, China

***Abstract* — In the field of health insurance, information technology and standardization of medical insurance are important area of work. They depend on each other in the medical insurance system construction and management service. They also form the foundation of building better information networks with the use of data mining. This paper describes the combination of K-means clustering technology and Naïve Bayesian classification technology to achieve this. The accuracy of data classification of health care costs can be enhanced to provide technical support for improving medical insurance with the benefit of an overall review of medical budgets and intelligent regulation. In this way, the efficiency of the medical cost audit is improved, thus reducing the waste of medical resources.**

Keywords - Data Mining; Health Care Costs; Anomaly Detection

I. INTRODUCTION

In "Decision of the Central Committee of the Communist Party of China on Some Major Issues Concerning Comprehensively Deepening the Reform", it is put forward clearly that " it is required to further reform medical insurance payment methods and perfect the system of universal health care" [1]. In 2014, the key point of human resources and social security work in Zhejiang province further required to establish and perfect the medical insurance regulatory information platform, strengthen the overall review of the medical expenses and intelligent supervision, strictly control cost control so as to create "sunshine health" [2]. Information standardization is the foundation to establish perfect information network of medical treatment insurance and the intelligent health monitoring platform. Main goal of data mining technology is to transfer data through standardized processing to be information that computer can process, and use the information in decision or verification so as for knowledge accumulation. Once start using data mining technology, effective analysis can be conducted for a large amount of precipitation data and find out the hidden rules and patterns so as to promote the development of medicine. [3] Therefore, informatization and standardization of medical insurance data are two interdependent important works in the medical insurance system construction and management services in, which are also the basis of improving the quality of medical services and build perfect medical insurance information network [4].

II. RESEARCH PURPOSE

This study is to use data mining technology to establish a set of classification model to assist the health care center in medical treatment charge examination work. The purpose of the study is summarized briefly as follows:

(1) Combining with the Naive Bayesian classification algorithm (NB) and K-means clustering algorithm to develop data analysis model that more effective.

(2) Find the key factors of the medical expense review.

(3) Use the data mining technology developed this study; assist in healthcare center for medical expenses audit business.

III. RESEARCH METHODS

A. Data set establishment

As part of the research, firstly the required medical costs data subset is created, containing the data set associated with 1000 clinics. At the same time, these data are anonymous, identification information are removed, including patient's ID, name, gender, age, ward, bed number, etc. Then these costs are reviewed by experts, manually assign a label, including "normal" and "abnormal" state.

B. Model design

Health care costs review is divided into two aspects: the administrative examination and the professional examination. Professional review need to judge the rationality of related expenses by combining with patient medical records for treating behavior. But the current health care costs audit does not involve patient record upload work. Therefore, this study focuses on administrative review, that is, focus on the overall review of reimbursement of medical expenses. This study assumes that in a medical institution, medical expenses produced by different doctors to treat same diseases is close or meet certain rules. Overall system architecture is as below (See figure 1).

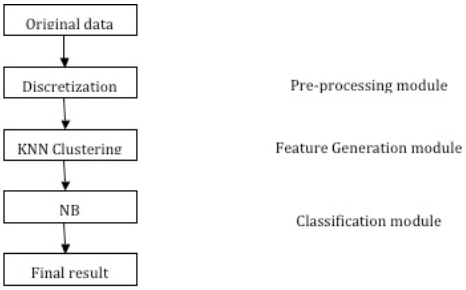

Figure 1 System architecture diagram.

C. *Establishment of the relevant indicators for the health care medical expense audit model*

Analyze the medical expenses audit business factors to form various initial properties in analysis model. Because it is assumed that all dimensions are independent each other in NB, we need to use correlation between each dimension and business rules to remove high correlation attribute to avoid the dependence between each dimension. For example, total amount = self-paying amount + reimbursement amount. Finally, the medical expenses of audit indicators are listed in the following table (see table I).

TABLE I MEDICAL EXPENSES AUDIT INDEX TABLE AND DATA DISTRIBUTION

	Dia.	*Ins.*	*Ope.*	*Med.*	*Ane.*	*Mat.*	*Dro.*	*Blo. Tra..*
Max	*960*	*14515*	*18542*	*162254*	*3660*	*48771*	*52617*	*4940*
MIN	*4*	*10*	*431*	*102*	*14*	*0.8*	*31*	*120*
MED	*356*	*350*	*2414*	*360*	*261*	*456*	*573*	*363.5*

D. *Preprocessing of the original data set*

Subjects are medical expenses summary data sets after anonymous treatment. It is conducted by following the standardized data mining step proposed by people like Hui. [5] Firstly, data need to be pre-processed, including data cleaning, merger, conversion [6], try to avoid incomplete, incorrect data impact on health care by audit data model. Data cleaning is for the purpose of collecting useful data set by performing a preliminary cleaning and filtering action to ensure data quality. This part include to check whether there is violation of the data of field constraint; check if there is null values in the clinic departments, such as diagnosis, health care category field; check if total cost is 0, etc. This process will remove data with something wrong from the business point of record, and check the integrity of the data record. Data consolidation is to establish data table in accordance with all relevant health care medical expense audit model required before. The purpose of data transformation is to ensure that the data format or type comply with the requirements of the data mining model and method. In view of this, first of all, dimension merger is conducted on the medical department, coarse-grained diagnosis in order to avoid these dimension values too scattered, which may influence subsequent generation based on K-means clustering feature. In addition, in the Naive Bayesian model, if the property value is continuous, the continuity of the numerical change will impact on the probability distribution enormously. At the same time, definition of the density function for processing dimension of continuity is complex. The continuity of data attributes need to first define the scope and the continuity of the numerical discretization, simplifying the calculation process and improving the calculation accuracy. In this study, the discretization method based on the information gain is used; determine the optimal set breakpoints through iteration to compare different breakpoint value of information gain value. Information gain (IG) [7] is defined as a information entropy difference between breakpoints on the information of s1 and subset S.

$$IG(s_1) = H(S) - H(s_1) \quad (1)$$

$$H(S) = -\sum_{i=1}^{n} s_i \log_2(S_i) \quad (2)$$

$$H(s_1) = \frac{|S_1|}{|S|} H(S_1) + \frac{|\overline{S_1}|}{|S|} H(\overline{S_1}) \quad (3)$$

Based on the above definition, discretization processing is conducted on diagnosis fee, inspection fee, and operation fee, expenses for medicine, anesthesia, materials, tests, blood transfusion, unpaid amount and reimbursement amount. After pretreatment, 996 effective data samples are gotten. The distribution of the data set is that 807 samples are normal and the "abnormal" tag samples are 189. From the two samples, 80% is chosen as training set and 20% for the test set.

E. *Generation of clustering features based on K-means*

K-means is a kind of semi-supervised algorithms that is widely used. [8] It is proved to be very effective when taken as supervision environment features, which can improve the performance of supervised learning algorithms.

In the training set $\{x^{(1)}, \ldots x^{(n)}\}$, each $x^{(i)} \in \square^{n}$ randomly selected cluster centroid point $\mu_1, \mu_2, \mu_3, \ldots \mu_k \in \square^{n}$, iterative process until getting the following convergence.

For each sample , calculate class it should belong to:

$$c^{(i)} = \arg\min_{j} \left\| x^{(i)} - \mu_j \right\|^2 \quad (4)$$

For each sample class j, recalculate the sample class of center of mass:

$$\mu_j = \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}} \quad (5)$$

Among them, $c^{(i)}$ is the closest category in samples i and k distance, $c^{(i)} \in \{1 \ldots k\}$. The decision of K is one of the most important decisions in cluster analysis. When K is big, the homogeneity of the classification is better, but will not

obtain meaningful classification methods; when K is small, although can better accomplish clustering, but with poor homogeneity. In this study, clustering characteristics of different medical departments, diagnosis, doctors of gender in different K - means periods are figured out for NB. Take $k \in \{2,4\}$ into the k-means for clustering analysis. Standard deviation produced by various clustering cluster evaluation and assessment decide the k value. The final standard deviation results are in the following table (see table II).

TABLE II DATA AGGREGATION CLASS STANDARD DEVIATION TABLE

	K=2	K=4
Data set (the 1st quarter)	*0.213*	*0.190*
Data set (the 2nd quarter)	*0.136*	*0.122*

Considering the final research goal is a binary classification, choose k to be 4.

F. Use NB to supervise data classification

In this module, use NB for the rest of the training set classification, distinguish between samples of normal and abnormal samples. [9] Through prior probability of a certain object, NB classifier calculate the probabilistic by the Bayesian formula [10], namely the probability that object belong to one kind, and select the classes with maximum a posteriori probability as the object's class.

Suppose: C is category node set, $c_j \in C$, $\langle v_1, v_2, \ldots v_n \rangle$ is attribute, and assume that condition of n feature attribute independent, then definition of $P(c_j \mid v_1, v_2, \ldots v_n)$ are as follows:

$$P(c_j \mid v_1, v_2 \ldots v_n) = \frac{P(v_1 \mid c_j) \times P(v_2 \mid c_j) \times \ldots \times P(v_n \mid c_j) \times P(c_j)}{P(v_1, v_2 \ldots v_n)} = \frac{\prod_{i=1}^{n} P(v_i \mid c_j) \times P(c_j)}{P(v_1, v_2 \ldots v_n)} \quad (6)$$

$$= \frac{P(c_j)}{P(v_1, v_2 \ldots v_n)} \prod_{i=1}^{n} \frac{P(c_j \mid v_i) \times P(v_i)}{P(c_j)} = P(c_j) \frac{\prod_{i=1}^{n} P(v_i)}{P(v_1, v_2 \ldots v_n)} \prod_{i=1}^{n} \frac{P(c_j \mid v_i)}{P(c_j)}$$

$$= P(c_j) \prod_{i=1}^{n} \frac{P(c_j \mid v_i)}{P(c_j)}$$

Definition: set N_{ex} as sample, N_{cl} as the number of classification nodes, and $N(c_j)$ belong to the classification of the sample c_j.

Laplace calibration:

$$P(c_j) = \frac{N(c_j) + 1}{N_{ex} + N_{cl}} \quad (7)$$

M estimates:

$$P(c_j \mid v_i) = \frac{N(c_j \& v_i) + m \times P(c_j)}{N(v_i) + m} \quad (8)$$

Among them, $N(v_i)$ is the quantity of samples that comply to v_i. $N(c_j \& v_i)$ is the quantity of samples that belongs to c_j and meet v_i. M is the correction coefficient, m = 2 in this experiment.

Therefore, we extend on the basis of the original EHRS Combined Classification Application (EHRCCA) [11], The NB implementation in the WEKA is added [12]. At the same time, the state of medical expenses audit prior probability of each binary Classification is calculated. Based on module produced before, a NB classifier model is set up by using discrete feature subset from the training set, and K - means clustering feature subset.

G. Post-processing module

Through processing of four components, each item of the training focus was allocated to two probabilities, respectively the probability belonging to "normal" and "abnormal". In this module, using the method based on rules, to calculate the final sample items described in the category of the state. Rules are defined as follows:

$$\text{Max}\left(P(c_1 \mid v_1, v_2 \ldots v_n), P(c_2 \mid v_1, v_2 \ldots v_n)\right) \Rightarrow Label_a \quad (9)$$

Among them, c_1 said normal classification, c_2 said anomaly classification, a said data entry of 1 sample in training focus article.

IV. PERFORMANCE EVALUATION METHODS AND RESULTS

We use general evaluation index in the field of medical statistics [13] on system performance and compared it with the performance-frequency distribution method. These indicators include precision, recall rate and F-score [14].

$$\text{precision} = \frac{\text{TP}}{(\text{TP}+\text{FP})} \quad (10)$$

$$\text{recall} = \frac{\text{FP}}{(\text{TP}+\text{FN})} \quad (11)$$

$$\text{F-score} = \frac{(2 \times \text{recall} \times \text{precision})}{(\text{recall}+\text{precision})} \quad (12)$$

In the formula, TP said gold standard result is normal; the actual classification results are normal sample size. TN said gold standard result is abnormal; the actual number of sample classification result is abnormal. FP said gold standard result is abnormal; the actual classification result is normal sample size. FN said gold standard result is normal; the actual number of sample classification result is abnormal.

Table III shows the confusion matrix on the test set of the output system [15]. Table IV shows the corresponding performance evaluation.

TABLE III CONFUSION MATRIX OF THE TEST SET OUTPUT ON THE SYSTEM

	Classification		
	Normal	**Abnormal**	**Total**
Normal	*142*	*19*	*161*
Abnormal	*6*	*32*	*38*
Total	*148*	*51*	*199*

TABLE IV SYSTEM PERFORMANCE EVALUATION RESULTS

	Classification	
	Normal	**Abnormal**
Sample Qty	*161*	*38*
Recall rate	*88.2%*	*84.2%*
Accuracy	*95.9%*	*62.7%*
F1	*0.919*	*0.719*

From the perspective of the evaluation results of the test set, F1 scores the classification of normal and abnormal classification of were 0.919 and 0.919 respectively, it can be seen that on the test set, model meet the design requirements.

V. CONCLUSION

This paper describes a data mining system that combines semi-supervised K-means clustering technology and naive Bayesian classification technology, which uses some of the common key indicators in the cost audit. Through the analysis of the health care costs data subset, and the discovery of hidden data we form the correct and fair audit model that can be used as reasonable supplement of existing mixed mode based cost audit and artificial sampling audit.

ACKNOWLEDGMENT

This research was supported by Medical and Health Planning Project of China (No. WKJ-ZJ-1518) from National Health and Family Planning Commission and Medical and Health Planning Project of Zhejiang Province of China (No. 2013KYB140).

REFERENCES

[1] "Decision of the Central Committee of the Communist Party of China on Some Major Issues Concerning Comprehensively Deepening the Reform" [OL].

[2] "Key Point of Human Resources and Social Security Work in Zhejiang Province" [OL].

[3] Hripcsak G, Bloomrosen M, Flatelybrennan P, et al. "Health data use, stewardship, and governance: ongoing gaps and challenges: a report from AMIA's 2012 Health Policy Meeting", J Am Med Inform Assoc, vol. 21, No. 2, pp. 204-211, 2014.

[4] ROSE J S, FISCH B J, HOGAN W R, et al. "Common medical terminology comes of age, Part One: Standard language improves healthcare quality", J Healthc Inf Manag, vol. 15, No. 3, pp. 307-318, 2001.

[5] Yang H, Spasic I, Keane J A, et al. "A text mining approach to the prediction of disease status from clinical discharge summaries", J Am Med Inform Assoc, vol. 16, No. 4, pp. 586-600, 2009

[6] Chapman W W, Nadkarni P M, Hirschman L, et al. "Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions", Journal of the American Medical Informatics Association, vol. 18, No. 5, pp. 540-543, 2011.

[7] FORMAN G. "An extensive empirical study of feature selection metrics for text classification", Journal Of Machine Learning Research, vol. 3, No. 3, pp. 1289-1305, 2003.

[8] Domingos P. "A few useful things to know about machine learning", Communications Of The Acm, vol. 55, No. 10, pp. 78-87, 2012

[9] Nadkarni P M, Ohno-Machado L, Chapman W W. "Natural language processing: an introduction", Journal of the American Medical Informatics Association, vol. 18, No. 5, pp. 544-551, 2011.

[10] Bastanlar Y, Ozuysal M. "Introduction to machine learning", Methods Mol Biol, vol. 1107, No. 1, pp. 105-128, 2014.

[11] Liang J, Zheng X, XU M, et al. "A combined classification model for Chinese clinical notes", International Journal of Applied Mathematics and Statistics, vol. 49, No. 19, pp. 201-209, 2013.

[12] Hall M, Frank E, Holmes G, et al. "The WEKA data mining software: an update", SIGKDD Explor Newsl, vol. 11, No. 1, pp. 10-18, 2009.

[13] Hripcsak G, Rothschild A S. "Agreement, the f-measure, and reliability in information retrieval", J Am Med Inform Assoc, vol. 12, No. 3, pp. 296-298, 2005.

[14] Yang Y. "An evaluation of statistical approaches to text "categorization , Information retrieval, vol. 1, No. 1, pp. 69-90, 1999.

[15] Uzuner O, Solti I, Cadag E. "Extracting medication information from clinical text", J Am Med Inform Assoc, vol. 17, No. 5, pp. 514-518, 2010.