International Journal of Engineering and Technology Innovation, vol. 2, no. 2, 2012, pp. 126-137

Use of Data Mining Techniques to Detect Medical Fraud in Health Insurance

Kuo-Chung Lin [1,2*], Ching-Long Yeh [2]

[1] New Taipei City Hospital, New Taipei City, Taiwan, Taiwan, ROC.

[2] Department of Computer Science and Engineering, Tatung University, Taipei, Taiwan, ROC.

Received 25 December 2011; received in revised form 26 January 2012; accepted 26 February 2012

Abstract

The health insurance claims application case the inspection usually relies on experts' experience for verification and experienced personnel in charge for checking. However, due to the heavy work load and the insufficiency of manpower and experience, the ratio of miscarriages of justice is high, leading to improper settlement of claims and the waste of social resources. This paper takes advantage of data-mining technology to design models and find out cases requiring for manual inspection so as to save time and manpower. Six models are designed in this paper. By the analysis of the 20/80 principle and the coverage and accuracy ratio, a great number of periodic data (over 2 million records) are fed back to the data-mining models after repetitive verification. Also, it is discovered that to integrate the data-mining technology and feed back to different business stages so as to establish early warning system will be an important topic for the health insurance system in hospital's EMR in the future. Meanwhile, as the information acquired by data-mining needs to be stored and the traditional database technology has limitations. Next time, this paper explores the ontology framework to be set up by semantic network technology in the future in order to assist the storage of knowledge gained by data-mining.

Keywords: data mining, health insurance claims, medical fraud cases

1. Introduction

The current high-frequency of insurance fraud events, the government medical insurance agencies suffered huge financial losses. In order to reduce the loss, the assessment that associated knowledge-intensive and complex nature, require a lot of human resources costs. We are hoping to use these characteristics in the development of insurance fraud reasoning system. These characteristics will be reviewed with an experienced expert assess of the rules of the medical insurance fraud prevention knowledge, for comparison. The medical insurance fraud characteristics include: Damage level insufficient information, suspected diagnosis of proof, insured low willingness to cooperate and Cause of the accident unreasonable [19]. Repeatedly claims record, in a special area, occur at a specific time and claims for late filing [15]. Inconsistent documents of application, high claims payments, certificate of poor reliability, non-cooperation and very familiar with insurance knowledge [4].

Public health insurance covers five categories: injury, disability, childbearing, death and agedness. Among them, the

* Corresponding author. E-mail address: ao5714@ntpc.gov.tw
Tel.: 02-29829111#3261; Fax: 02-29884964

injury payment is classified into two categories which include ordinary accident health insurance and occupational accident health insurance that is used to indemnify laborers for occupational accidents. To judge the injury caused by occupational accidents, experience and knowledge required to involve many professional fields. Thus, only if connection and integration are made among various fields, can the judgment capability be achieved to carry out claim cases. Currently, the procedures to deal with claims for occupational accident health insurance mostly rely on the experts to manually conduct the following procedures of "receiving-inspection-payment". (1) Receiving: Integrate the experience and relevant applications noted by professors and conduct preliminary screening of the cases. (2) Inspection: Judge and select potentially questionable cases with the experience of receiving units and personnel and review for the second time. (3) Payment: Carry out on-site review or further review of the written documents in regard to the cases screened out; then inspect the settlement results and make payments on the basis of the review results. As medical fraud cases to falsely claim the health insurance money emerge one after another, auditing and screening of those cases should be strengthened. However, due to the heavy workload and limitations of human resources and experience, it is not easy to screen all the cases one by one carefully. Besides, during the process of judgment in every step (receiving, inspection and payment), the experience cannot be shared easily. Even though education and training have been given, the result is still far from satisfaction. Obviously, the above mentioned problems cannot be well solved by wholly depending on manual operation. Therefore, this research turns to information technology for the promotion of the degree of automatic processing so as to solve the aforesaid problems.

At present, the way to screen questionable cases is to receive all the cases first and to arrange manpower to check through the information system one by one whether the application data are in consistency. If the data are consistent, check by manual comparison whether the hospitals and doctors issuing the certificates are registered medical institutions and doctors. Then, by comparison with historical cases and records of the insurant, check whether the case is questionable. As the health insurance is national wide, the quantity of cases is so large that the process costs forty-five days. If it is judged to be a questionable case, courts in all places should be noticed in written to assist the investigation. It takes nearly thirty days to identify whether it is a medical fraud case or not. Thus, the whole process needs about seventy-five days, costing a huge man-hour.

The method of this research is to solve the problem of manpower input and experience by the utilization of data-mining technology. First, according to the 20/80 principle, it is expected to screen the largest percentage of questionable case with the smallest sample size under fewer costs. In another words, it is attempted to cover 80% of questionable cases with the sample size of 20% so as to decrease manpower input for review. Secondly, relevant rule of thumb is to be found out in the process of data-mining for the reference of inspectors and reviewer. The third step is to integrate data gained by data mining into the processing stages (such as receiving, inspection and payment). Medical fraud cases can be found out as early as possible with the help of the effective marks of the application system and the efficiency of audit and review can be enhanced, and also preliminary research can be carried out for the study of the future ontology knowledge management system.

The research expects to adopt the approach of data-mining to effectively cover a large number of cases possibly arising among the data at the expense of a few data and research costs. In this way, human resources and costs consumed due to the heavy work of manual inspection can be reduced. Meanwhile, cases which require experts to review previously can now be fulfilled with the help of knowledge reservation and transfer, realizing the promotion of the core competency. The second part of this paper explains the identification of public health insurance claims and problems to be confronted with. Also, it discusses the procedure of knowledge discovered by data-mining. The third part demonstrates the manner of research design, data checking and develops models. The fourth part trains the models with practical data. In the fifth part, the models are adjusted based on the results acquired. The basis and method of adjustment are summarized. In the final part, discussions are put

forward in regard to the research results and recommendations. Also, the direction of application in the future is indicated.

2. Literature Review

2.1. Identification of Health insurance Claims

In health insurance, the occupational accident health insurance is the most important. However, it is difficult to determine the meaning and scope of an occupational accident. Owing to the difficulty in deciding which accidents can be deemed as occupational accidents, disputes have been caused in practices and in the academic circle [3]. An occupational accident refers to the fact that the normal process of production technology is hindered by abnormal phenomena caused by the deficiency of production tools and means. Some paper to define occupational accidents as “accidents of body injuries, which refer to incidents unexpected or out of control that incur harms to human body because of objects, materials, and people or radiation effects [11]. The Council of Labor Affairs has made plenty of laws and regulations on the identification of occupational accidents, among which it is provided in Item 4 of Article 2 of Labor Safety and Health Act that “the so-called occupational accidents involve diseases, injuries, disabilities or death of labours caused by architecture, equipment, raw materials, stuff, chemicals, gas, steam, dust, etc. in the laborers’ working areas as well as other occupational reasons” [6].

In accordance with the Labor Health insurance Act, the laborers health insurance is divided into two categories: (a) Ordinary accident health insurance: childbearing, injuries and sickness, medical treatment, disability, unemployment, agedness and death. (b) Occupational accident health insurance: four kinds of payment and disappearance allowance—injuries and sickness, medical treatment, disability and death [5].

From the above literatures, it can be known that to identify an occupational accident need to conduct a very complicated decision procedure of claims. It cannot be easily settled only by information system but requires legal experts or professional personnel familiar with related aspects, and time and costs are also problems. In this research, the case institution not only provides relevant health insurance of occupational safety for people, but also plays a role in promoting the sense of identity of the public towards the government. Besides undertaking various health insurance businesses, it also deals with accounting and payment of different claims. The process is quite complicated in essence and the decision procedure of different claims adds to the complexity. Therefore, how to save labor costs in virtue of the information technology and to precisely find out abnormal cases will be a significant subject.

2.2. Data Mining

Data mining can to discover new facts, rules or relations from the raw data [9]. Data mining is a process, one of the procedures of knowledge discovery, to discover useful models in the data probably to be used in the future, which have never been seen previously and are easily to be understood [18]. Data mining is the analysis of data in an automatic or semi-automatic way, expecting to discover meaningful rules or knowledge from it [21]. Data mining is to find out information with special meanings from a great number of data by some technology as the procedure to discover knowledge from the database, the steps are as follows [13]: (a) Data cleaning: remove noise data and inconsistent data. (b) Data integration: combine various data sources. (c) Data selection: find out data related to the subject from the database. (d)Data transformation: transform data into the form appropriate for mining. (e) Data mining: extract data models by the utilization of technology. (f)Pattern evaluation: evaluate models really useful to present knowledge. (g) Knowledge presentation: present knowledge after mining to the users by using technology such as visual presentation. Therefore, data mining is to adopt the mining procedure to discover unknown knowledge and rules from plentiful data. To sum up, this research employs the method and steps as the data-mining process and adopts SPSS Clementine 7.1 as the tool to establish and adjust models.

2.3. Insurance Fraud Detection

In this section, we discuss the academic to identify characteristics of the insurance fraud case. Hope to use these characteristics in the development of insurance fraud reasoning system. These characteristics will be reviewed with an experienced expert assess of the rules of the insurance fraud prevention knowledge, for comparison. Many academics have proposed ways to prevent fraud. For example, using statistical techniques to explore the incidence of fraud cycle [1]. The establish Private Network of Social of control investigative insurance company [10]. This survey describes the range of these moral hazards arising from asymmetric information, especially in claiming behaviour, and the steps taken to model the process and enhance detection and deterrence of fraud in its widest sense [8]. In automobile insurance fraud prevention studies case, it will assess the results distinguish between two kinds of results, suspected fraud cases and reasonable. Prediction method use Logistic Regression [17]. Other studies want to classify for different levels, use of Fuzzy C-means, Artificial Neural Network and Line Regression [19] [7]. Some scholars in the life insurance claims fraud, the use of game theory to infer information asymmetry caused by fraud factor [16] or use the least square method to identify the factors that affect the insurance fraud cases can apply to increase the recognition rate [12].

3. Research Design

3.1. About "Health insurance"

Health insurance for the Taiwan region is responsible for workers, farmers, fishermen receive government health insurance agencies, mainly for the provision of employment of health insurance process, the scope of its benefits by the insurer containing the reproductive benefit, injury benefit, disability benefit, old-age benefits, death benefits, such as missing payments-related businesses, mainly in the protection by insurers and in the event of emergency, through the social welfare system to help the insurer to maintain a normal life with their families. This study through the scientific technology to assist the original artificial case of the admissibility of the above applications, and to prevent medical fraud as a result of losses caused by the community, thus putting in a lot of review about 700,000 cases per a month on the human cost of the required cost and time. Therefore, the effective use of scientific method in the unrelated tradition of the original database to find information is the focus of this study.

3.2. Method and Steps

The business process of health insurance claims is indicated in Fig. 1. If the insurant has an accident, after being treated in medical institutions, he/she should fill up the application form in the unit which he/she belongs to and submit relevant supporting documents to the health insurance company. Upon receipt, data should be entered in the captive health insurance application system. Then, payment decision will be made based on the contents of the examination mechanism and the retrieval of relevant medical records. After the decision is made, actions will be taken accordingly to the claims.

The steps of this research method can be divided into four stages: data check, variable analysis, model development and model adjustment. Main work in every stage is as below: (a) Stage 1: Include interviews of sources of problems and requirements, the identification of problems, the fitness of raw data and relevant data collection as well as data pre-processing and data audit if necessary. (b) Stage 2: Select an available variable on the basis of information collected in Stage 1; analyze whether the variable needs to be transformed and delete data of great deviation after the retrieval and summarization of the procedure. (c) Stage 3: Construct and adjust the model; conduct the matrix analysis and cross validation with the 20/80 principle to determine the relevancy between the variable and the model and construct the mining model. (d) Stage 4: Select the best model pursuant to the 20/80 principle and the percentage of the coverage and accuracy ratio; adjust the model with the knowledge of the business process to establish the model.

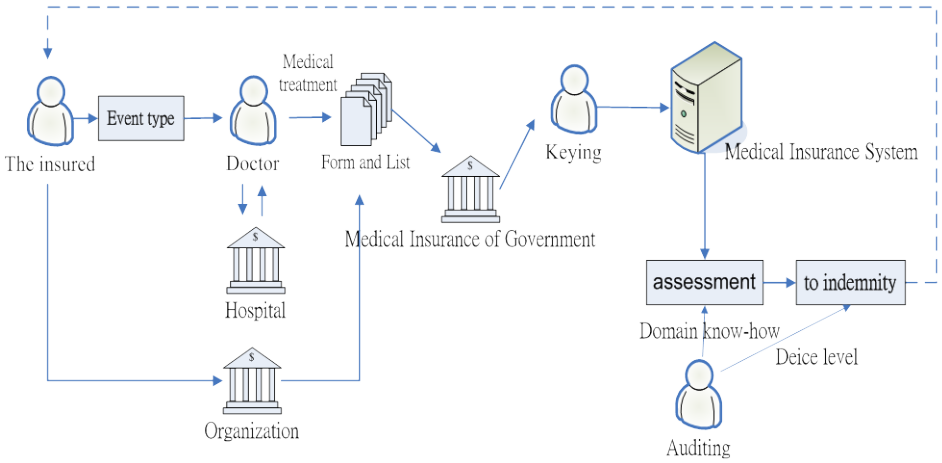

Fig. 1 Business Process Diagram of Health insurance Claims

3.3. Steps of Model Development

There are three steps for model constructing. The work step is shown in Fig. 3.

Step 1: By the combination of data sources documents required by the models, screen the target variable Medical fraud (Y); establish rules for data transformation in the meantime. (1)

Step 2: Based on the time points of the samples, select data required (as is indicated in Fig. 2-3) and construct data of training group and verification group. The training group data is used to develop the decision tree model while the verification group is to evaluate the actual performance of the model. (2)

Step 3: Adopt related algorithm to classify and structuralize the training group data. According to data grouped by the algorithm (decision tree), take advantage of the verification group data to transform the decision tree and obtain relevant information about which principles will make the target variable true. (3)

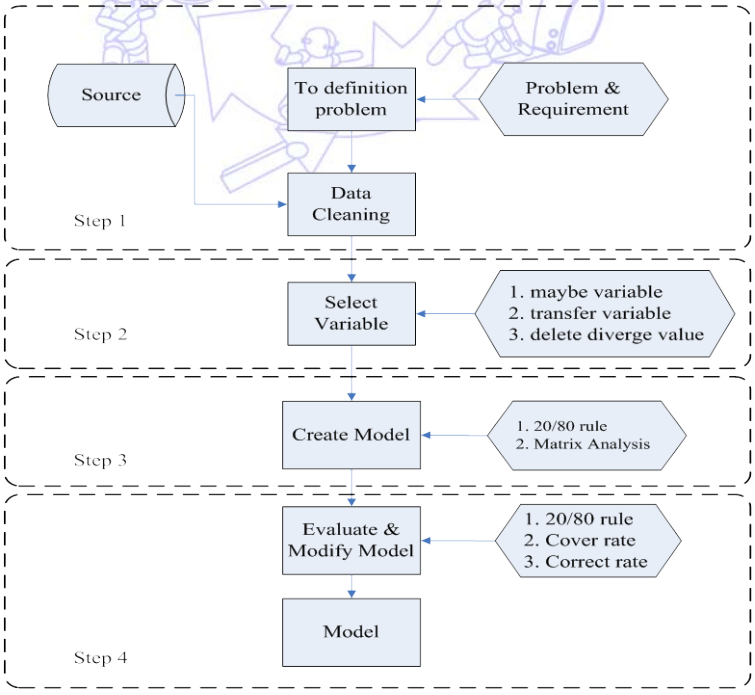

Fig. 2 Four Stages to Establish the Data-mining Model

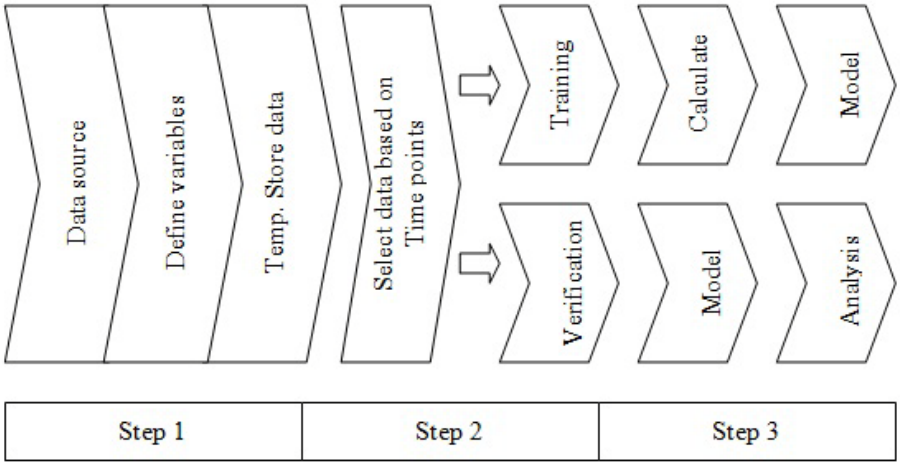

Fig. 3 Steps to Establish the Data-mining Model

As historical data can be used to verify the data intervals, at first, the preliminary model can be designed as required by the training group data as well as the classification of these data in terms of related algorithm, so as to infer the accuracy and coverage ratio. In the research, it is expected that the coverage ratio is above 70% while the accuracy ratio of over 60% can be obtained.

3.4. Data-Mining Model

Follow in chapter 2, Characteristic the integration of medical insurance fraud as follows:

(1) Analysis of high claims payments: for example, too many hospital days, more than the standard 180 days, the claim amount is too large, consistent with features of suspected insurance fraud. Need to control two variables:

Days of hospital stay: Use insurance number (IDNO), the history of the insurance payment for the number of days hospitalized for file (BMSF2) and cumulative disability hospital file (BMSF3) the cumulative length of stay plus the total (CHKDAY = A & B). The total combined file this application (BBDET) of hospital stay CHKDAY = C. If A + B + C => 180, the claim amount is too large for suspicious characteristics of the insurance fraud. (Model 1)

Insurance seniority: Length of coverage claims payments, the higher base, such as insurance years less than 1 year, compared with cases of suspicious applications. From the insurance number (IDNO), check the capital personal data files (DWPT) of the initial application date (RRDTE) with this application file (BBDET) the date of application (EVTDTA) differences. (Model 2)

(2) Data is incorrect: if the applicant insurance information (including insurance eligibility, insurance compensation, the insured amount) in the transaction within 30 days, may meet the eligibility criteria for applications for change of personal data, for the suspicious behavior of the insurance fraud.

Validity of the insurance status: through the Insurance number (IDNO), the insured unit file (DWPB), to obtain insurance unit code (UBNO), check the status of the insured (QMK) is valid. (Model 3)

Transaction data deliberately: In order to check the data on stability, personal capital in the data files (DWPT) to find information on transaction date (EFDATA) - The date of application (EVTDTA), need more than 30 days. (Model 4)

(3) Suspicious data analysis: high percentage of salary adjustment, the ratio is greater than 30%, may deliberately raise money to prepare insurance claims. No fixed work, and through professional associations to join the insurance, so reliability is a shortage of information for possible insurance fraud.

High percentage of salary adjustment: the insured unit (DWPB) salary (WAGE) and the insured (DWPT) historical average salary (AVWAGE) ratio higher than 30%, was suspicious of the insured salary. (Model 5)

(4) Problems of hospitals or physicians: for the insured had provided false proof of hospitals or physicians. Government medical insurance has their control, than on the blacklist file so it is important to the overall review of the insurance fraud aspect. Rules are as follows:

Suspicious hospital: Use security number (IDNO) to the certificate file (BBDET) in the corresponding hospital code (EVIDNO), and check the notes (MRK-H) field, not a "1" Note. (Model 6)

The above analysis and data mining needs of combination, the data mining required integrating data sources in Table 1.

Table 1 Data mining models and data source analysis

Characteristic of Insurance fraud	Data Mining Model	Source Table	Colum
High claims payments	Insurance fraud relations with Days of hospital stay	BBDET	IDNO, CHKDAY
		DWPT	IDNO
		BMSF2	CHKDAY
		BMSF3	CHKDAY
	Insurance fraud and Insurance seniority relations	BBDET	IDNO
		DWPT	IDNO, RRDTE
		BBDET	EVTDTA
Data is incorrect	Insurance fraud and Validity of the insurance status relations	BBDET	IDNO
		DWPT	IDNO
		DWPB	UBNO, UNTYPE, QMK
	Insurance fraud and Transaction data deliberately relations	BBDET	IDNO
		DWPT	IDNO, EFDATA
		BBDET	EVTDTA
Suspicious data analysis	Insurance fraud and the Reliability of the insured organization relations	BBDET	IDNO
		DWPB	IDNO, UBNO, UNTYPE
	Insurance fraud and High percentage of salary adjustment relations	BBDET	IDNO,
		DWPB	IDNO, UBNO, WAGE
		DWPT	IDNO, AVWAGE
Problems of hospitals or physicians	Insurance Fraud and Suspicious hospital relations	BBDET	IDNO
		BBDET	MRK-H
	Insurance Fraud and Suspicious of physicians relations	BBDET	IDNO
		BBDET	MRK-D

4. Model Training

In this research, discuss Data-Mining Source and six models adopted have different mining purposes, data intervals and samples and employ 20/80 Principle analysis with Gain Chart and matrix analysis for the coverage and accuracy ratio. Here analysis and explanation are given in respect of Model 1 and Model 4.

4.1. By model first running results

Model first running of the data output results shown in Table 2.

Table 2 by model first running results

Item	Count (P)	Y=1 (A)	Rate (A/P)	$R-Y1 (B)	Covered (C)	Accuracy Rate (C / B)	Covered Rate (C / A)
Model 1	19562	487	2.5%	948	393	41.46%	80.70%
Model 2	5002	479	9.6%	381	235	61.7%	49.1%
Model 3	4939	226	4.6%	293	184	62.8%	81.4%
Model 4	79	27	34.2%	32	26	81.25%	96.30%
Model 5	3914	698	17.9%	739	592	80.1%	84.8%
Model 6	7782	414	5.3%	1982	377	19.0%	91.0%

4.2. 20/80 Principle analysis with Gain Chart

In terms of Gains Chart of Model 1, which indicates the relevancy of the model from the perspective of the 20/80 principle, Fig. 4 is not the optimal ideal condition as nearly 35% must be mined to find out nearly 80% of cases. With regard to Gains Chart of Model 4, the condition in Fig. 8 requires to mine nearly 35% to find out 80% of cases. This represents that from the initial mining, it is discovered that the model must be adjusted so as to achieve better effects.

4.3. Matrix Analysis

In Fig. 4 of Model 1, the parent matrix of the data sample is 19562 and claims which have actually happened are 487. Among the 80.7% of the coverage ratio, the accuracy ratio is above 41.6%, which is a low result. In Fig. 5 of Model 4, the parent matrix of the data sample in this interval is such a small number as 79 while the actual claim cases are 27. However, among the coverage ratio 96.2%, the accuracy ratio is above 81.3%. Although the coverage and accuracy ratio in Model 4 is good, it is quite a pity that the parent matrix is too low.

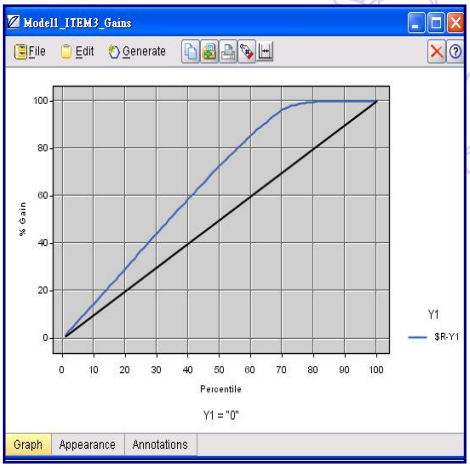

Fig. 4 Model 1 Gains Chart

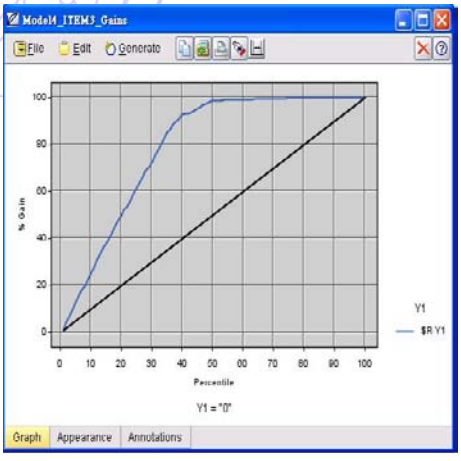

Fig. 5 Model 4 Gains Chart

4.4. Meta-analysis

The results of the models acquired through data-mining in the first stage are indicated in Table 2. The accuracy and coverage ratio of Model 1 is not so satisfied that the model should be adjusted in order to cover more and be more accurate. In this way, the exploration of cases can be conducted effectively by using a small amount of data.

5. Model Adjustment

5.1. Reduction of Processing Data Sets and Vacant Data

In order to improve the coverage and accuracy ratio of the model, the first adjustment mainly aims to reduce the processing of data sets and vacant data. Adjustment will be made to the model by adopting data intervals of Stage 3 and 4. The adjustment method and actions are as follows: (a) Select key fields of tables whose value is more than one million for trim. (b) Set a temporary storage location for the nodes of the model to reduce the load of the system memory. (c) Adjust the maximum of data sets to be 20000 for fear that the number of data sets is too large. (d) Define blanks in the model nodes. (e) Fill up the values of blanks by the Filler node. (f) Add a new Filter node to block fields that do not need to be entered in the machine for study. (g) Control the maximum sample number to ensure smooth performance.

5.2. Preprocessing of Removing Data Repetitiveness and Increasing Data Representativeness

(a) Removal of data repetitiveness: Find out problems of checking or repetitive claims from the primary value of the table. In order to avoid that the combination of every repetitive data with other tables lead to an abnormally inflated total number, correction has been made and the difference from the previous version is indicated in the red frame. (b) Increase of data unique representativeness: After data processing, representative problems will emerge. For instance, if the first ten numbers fail to represent a unique identity card, its combination with other tables will result in an abnormal total number. Therefore, relevant preprocessing is added. (c) Considerations of the improvement of equipment performance. (d) Data should be stored in other places to avoid repetitive extraction from the enormous table documents and influencing other people's operation. (e) Due to the limited space of the hard disk, unnecessary data for executing the model should be eliminated as many as possible for fear that the temporary storage disk requires too much space.

5.3. Verification

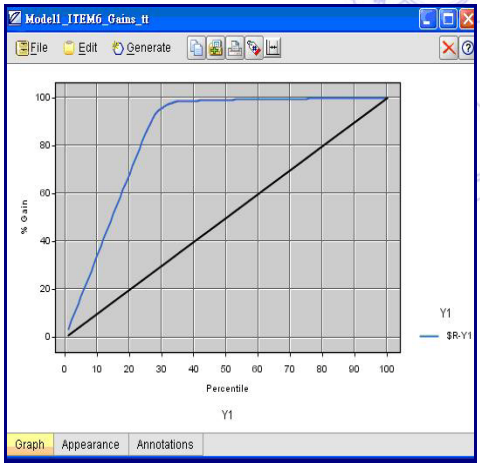

Fig. 6 Model 1Gains Chart after Adjustment

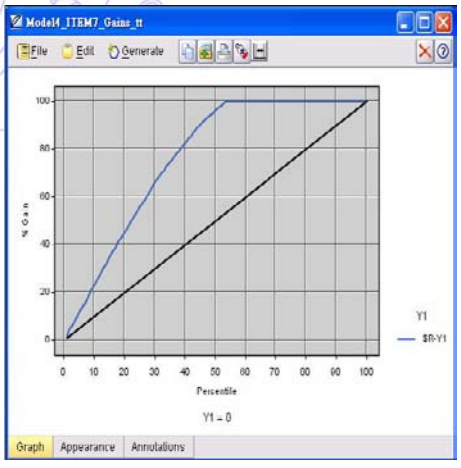

Fig. 7 Model 4 Gains Chart after Adjustment

Table 3 Accuracy and Coverage Ratio after Model Adjustment (%)

	Parent Matrix (P)	Y=1 (A)	Ratio (A/P)	$R-Y1 (B)	Covered (C)	Accuracy Ratio (C / B)	Coverage Ratio (C / A)
Model 1	17041	419	2.45%	527	381	72.30%	90.93%
Model 4	103	49	47.57%	49	37	75.51%	75.51%

Table 3 indicates the hierarchy analysis of the reliability of every model as well as the summarization of related icons and verification data. The output result of Model 1 is that it covers 90.93% of cases while the accuracy ratio is below 72.30%; that of Model 4 is that 75.51% of cases are covered with the accuracy ratio of below 75.51%. The results of the two models are

much better than that before the adjustment, which proves that by the manual screening of a few data the two models can achieve 70% of the accuracy of wholly manual operation previously. The confusion matrixes of these two models are shown respectively in Fig.6 and Fig.7.

5.4. Adjustment Conclusions

The research discovers that the following practices is helpful to adjust the accuracy and coverage ratio of the model, which include the removal of unnecessary data, making data more representative and the preprocessing of missing data. (a) Removal of unnecessary data: The pattern of question inquiry or thinking can be changed so as to remove needless samples. For example, compared with all applicants or people with or without payment, the relative minority can be removed preferentially when the model is constructed. (b) Data representative: Where appropriate, advanced data processing methods can be added (such as existing data generated from some statistics) to reinforce the rationality of data interpretation and summarization and to make the data abundant. (c) Processing of data missing: The missing of some data in the database will bring about difficulties in analyzing. Although some methods can be adopted to supplement the data, the data are not all real and may be inaccurate at times. Therefore, sometimes rules for removing or methods and principles for supplement the missing data can be thought about.

6. Research Conclusions

6.1. Research Findings

From Table 2, though Model 6 may need readjustment as its coverage and accuracy ratio are not sufficient. In respect of the interval data, if the models proposed in the research are adopted, the accuracy of random inspection can be promoted and the labor costs for complete checking can be saved as well.

(1) Feasibility of the application of the 20/80 principle to data mining

The research finds that it can really reduce the labor input in the review to screen the largest percentage of questionable cases with the smallest sample size by applying the 20/80 principle (for instance, 20% of the sample size can cover about 80% of the questionable cases). Meanwhile, the original dependence on the high input of labor costs can be relieved, which is the biggest benefit of the research.

(2) The technology and knowledge transfer needs to be broken through

It seems that manual operation is inevitable. Thus, how to transfer knowledge should be considered further for a solution. Currently, experts are employed or experience and knowledge of inspectors are relied upon. If the method can be found out to share the experience of the receiving units with the inspection units and of the inspection units with the payment units, better solutions can be provided for latest inspectors.

(3) The speed of feedback mechanism for finding out medical fraud cases

If the mining results can be integrated with the application system of every business stages (such operation stages as receiving, inspection and payment etc.), medical fraud cases can be found out early with the help of effective marks of the application system and the performance of audit and inspection can be enhanced.

(4) It is obviously insufficient to use traditional databases to store information gained by data mining.

From the research, it can be found that information acquired by data mining is different in type of storage from databases

of the ordinary application system. The traditional data type cannot satisfy the requirement. Besides, if data-mining should be carried out among a batch of considerable transaction data, stable models can be constructed to provide a tool for regular influx of the mass data and the data acquired can also be stored in a new kind of knowledge base in semantic knowledge type. This finding appears to be identical to the development requirement of Web 2.0 have 7 Step in the ontology construction proposed [6].

6.2. Research Recommendations

(1) Utilization beyond data and models required by data mining

Besides data required by traditional data mining, the two points below can be considered as well. (a) Non-disaggregated models can be taken into consideration. Although probably these can hardly be predicted, the intensity of relations between each other can be understood and even the cause and effect relationship can be inferred. (2) Safety and data grading mechanisms can be increased to provide the application field for data mining.

(2) The official organs should develop the data mining combined with the application system.

The science of data mining has been widespread. Yet, it should be developed how to feed back the data-mining contents to the application systems of every official organ by adopting notes of weighting or other methods. Meanwhile, this can also exempt the ordinary researchers from the embarrassment of mining only for data-mining's sake and increase the benefits of data mining.

6.3. Limitations and Future

Taking account of the confidentiality of businesses in official organs, the research keeps the contents of relevant data as secrets. However, the applications of data mining are not influenced thereof. Moreover, the practical application of this research consists in the cases applying for payment of injuries and diseases as well as disabilities. As all kinds of health insurance all have their respective professional fields, the mining and summarization of rules cannot be implemented comprehensively. These are the limitations of this research. Data gained by data mining, which contains plentiful information, may need a precise knowledge management system with meta-data management technology. Besides, the development technology of ontology and the future semantic network of Web 2.0 may solve existing problems. Therefore, the future research will aim at the development of a platform which is unable to store data gained by data mining.

References

[1] Bolton, Richard J. and David J. Hand., "Statistical fraud detection: a review," Statistical Science, vol.17.3, pp. 235-249, 2002.
[2] C. F. Lee, "Aspects of data mining," Information and Education, pp.10-19, Feb. 2001.
[3] C. G. Hwang, R.O.C.'s Labor Standards Law, National Open University, Taipei County, Mar. 2002.
[4] Clark, J., T. Davies and H. Tilley, "Fraud Investigation: A Claims Handler's Guide," http://www.crawfordandcompany.com/pdf/fraud, October 24, 2003.
[5] Council of Labor Affairs, Table of Actual Premium Changes Applicable for Occupational Accident Health insurance of Labor Health insurance, Labor Health insurance, 2006.
[6] Council of Labor Affairs, Table of Business Category and Premium Applicable for the Occupational Accident, Labor Health insurance, 2005.
[7] Derrig, R.A. and Ostaszewski, K., "Fuzzy techniques of pattern recognition in risk and claim classification," Journal of Risk and Insurance, vol. 62, pp.447-482, 1995.
[8] Derrig R. A., "Insurance fraud," Journal of Risk and Insurance, vol. 69.3, pp. 271-287, 2002.

[9] Grupe F. H. and Owrang M. M., Information System Management, Minnesota: Management Information Systems Research Center, 1995.
[10] Ghezzi and Susan Guarino, "A private network of social control: insurance investigative units," Social Problems, vol. 30.5, pp. 521-531, 1983.
[11] Heinrich H. W., Peterson D. and Roos N., Industrial Accident Prevention, New York: McGraw-Hill, 1980.
[12] Robert E. Hoyt and Colquitt L. Lee, "An empirical analysis of the nature and cost of fraudulent life insurance claims," Insurance Regulation, vol. 15, pp. 451-480, 1997.
[13] Han J. and Kamber K., Data Mining: Concepts and Techniques, San Francisco: Morgan Kaufmann Publishers, 2001.
[14] M. Stone. J., The Royal Statistical Society, pp. 36, Feb. 1974.
[15] Manuel Art´is, Mercedes Ayuso & Montserrat Guill´ en, "Modeling different types of automobile insurance fraud behaviors in the spanish market," Insurance: Mathematics and Economics, vol. 24, pp. 67-81, 1999.
[16] Pierre Picard, "Auditing Claims in the Insurance Market with Fraud: The Credibility Issue," Journal of Public Economics, vol. 63, pp.27-56, 1996.
[17] Sharon Tennyson & Pau Salsas,"Patterns of Auditing in Markets with Fraud: Some Empirical Results from Automobile Insurance," Working Paper, 2000.
[18] Bureau of National Health insurance, "Codes and scopes of classifications of diseases," http://www.nhi.gov.tw/02hospital/hospital_5.htm, Aug. 2005.
[19] Weisberg, H.I. & Derrig, R.A., "AIB Claim Screening Experiment Final Report," Insurance Bureau o f Massachusetts, 1998.