ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(43), DOI: 10.17485/ijst/2016/v9i43/104971, November 2016* ISSN (Online) : 0974-5645

Distance based Model to Detect Healthcare Insurance Fraud within Unsupervised Database

Hojjat Ahmadinejad[1], Amir Norouzi[2*], Ahura Ahmadi[3] and Ali Yousefi[4]

[1]Engineer of Information Technology, NouAndish Pars Co., Tehran, Iran; Hojjat.ah21@gmail.com
[2]Health Management and Economics Research Center, Iran University of Medical Sciences, Tehran, Iran; Dr.a.norozi@gmail.com
[3]Shahid Beheshti University of Medical Sciences, School of Medical Education, Tehran, Iran; ahura.ah.gs21@gmail.com
[4]Engineer of Computer Engineering, Department of Computer Engineering, Malayer Branch, Islamic Azad University, Malayer, Iran; a.yousefi.6764@yahoo.com

Abstract

Objectives: Healthcare fraud costs the country tens of billions of dollars a year. **Methods:** Fraudulent behaviors of healthcare providers and patients have become a serious burden to insurance systems by bringing unnecessary costs. Insurance companies thus developed methods to identify fraud. **Results:** In this paper a methodology offered based on data mining approach to discover fraud in healthcare insurance. **Applications:** To test and evaluate model real-world data set related to healthcare insurance in Iran has been used. Investment result of operation model on this data set indicates proper performance of it.

Keywords: Anomaly Detection, Data Mining, Healthcare Fraud, Outlier Detection, Unsupervised Method

1. Introduction

Medical insurance is one of the most essential types of insurance and the most critical and the most basic needs of people in modern societies, this has been the cause due to population growth, the number of insured people has increased dramatically and providing healthcare to them have wide dimension. On the other hand, increasing population aging is becoming a challenge for healthcare system in all around the world[1]. Healthcare price increases in old age, and various governments faced this issue. In the process of health insurance natural and legal persons such as healthcare, medical and pharmaceutical, hospitals, clinics, laboratories, doctors, pharmacies and ..., on the one hand. The insured patients from other hand, health insurance offices and centers which carry out all operations related to insurance process participate as the third party of the process. In this process, as well as other matters, may jobbers to pursue their interests in seeking fraud and receiving treatment services and drug illegally. Fraud in insurance claims may be occurred in any of the process steps by any of the participants in the process. Medical insurance fraud is a serious threat to public funds and public trust[2]. In 2011, $2.27 trillion was spent on health care and more than four billion health insurance claims were processed in the United States. It is an undisputed reality that some of these health insurance claims are fraudulent. Although they constitute only a small fraction, those fraudulent claims carry a very high price tag[3]. Healthcare fraud costs the country tens of billions of dollars a year. It's a rising threat, with national health care expenditures estimated to exceed $3 trillion in 2014 and spending continuing to outpace inflation[4]. According to high cost that fraud impose to

* *Author for correspondence*

health insurance, discover such cases is one of the most important parts of organizations. In the meantime, fraud detection leads not only to reduce costs, but also can be effective to prevent a shortage of some medicines scarce and expensive, and also in order to improve service delivery through more efficient use of medical resources available[5]. Some research reported specific fraud scheme detection using data mining approaches[6]. Data mining is gaining more attention by researchers as a potential tool to find health care fraud more easily[7]. Most of the studies are considered outlier detection as one of the primary tools[8]. Since it's already in most large databases of people covered by insurance and services provided to them, but still system processes to detect fraud have not developed, in this paper aims to identify suspected cases among large volumes of stored data, and algorithms to provide a system that can be used off-line. The main features of this algorithm, first is that results are descriptive; means in identifying suspicious case, also offers the relations between relevant features. Thus, by connecting the off-line system to the on-line, on-line system also provides the possibility to update automatically rules. The second feature of this algorithm is simply set computing; so that the algorithm can be used with a regular PC running on a large set of data and extracted the right results in very little time. This paper is based on the idea that the presented, and to improve and develop the ideas presented in it[9]. In the study, Aral and colleagues a method based on distance using data mining has been provided that through pairs examining the different characteristics to assess the likelihood of fraud in the medical prescriptions. This model on comprehensive database of adult heart surgery in Turkey was examined, and the results were obtained TP = 77.4% and FP = 6%[9]. In extracted fraud indicators and rules based on experience and knowledge of insurance professionals to produce an expert system in order to perform insurance activities is produced. EFD, an expert system, which is produced by Travel-Insurance finely knowledge (for behavioral initiative) apply, along with information theory to infer rules for detecting fraud. Another way in research was formed, the recent focus on the use of new machine learning techniques, in which all features is considered by council experts, determinate and production design inference was used for machine learning[10]. In a study that was done in America by Sokol, a model to distinguish between fake and real demands was designed and built. For example, in the system for each health service such as laboratory services, radiology, physiotherapy and determined a set of characteristics and then produced an inference model on which the system will help in identifying unusual requests[11,12]. In separate studies carried out in Australia for health insurance committee, offered a model for detecting errors of providers of insurance services. In this model, after determining the distinctive features by experts an algorithm based on neural network and fuzzy logic has been used to build the model search. Similarly, in present Case management demands a system based on fuzzy logic for insurance service providers [m, n, o]. Similarly, in Case management demands a system based on fuzzy logic for insurance service providers presented[13-15]. In[16] introduced a healthcare application processing requests that of a natural language processing engine and using text mining techniques to uncover errors and potential fraud pays. In a process framework to detect fraud or errors done by health insurance has been introduced by service providers. In this study, the structural pattern mining methods to determine a set of structural patterns of clinical samples (clinical) have been used[17]. In a data mining system to detect fraud in scope that the number of participants and their relationships is a lot like medical insurance have provided. In this study, using a combination of learning methods, decision trees, logistic regression, weighted local, k-means clustering and regression a fixed model provided that a ranking of the companies involved in such processes based on the probability of fraudulent operations in the future based on current data supplied[18]. In provide a method for ranking the candidates in their prescriptions. In this study, of per domain a version of normal behavioral model designed and by statistical methods identifies the items which had the greatest deviation from the model. This method identifies more than abnormal and all identified cases were right[19].

2. Proposed Approach

Research in fraud discovery on supervised algorithm field usually focuses non-linear. But we need less complex, reliable and faster algorithms. Since our data here are without labels identifying fraud and legal cases, our approach will be used unsupervised. We need two tools to review medical transactions. One batch screening/ auditing as an off-line system and a system of on-line/

on time transaction control. So we must create these two systems that interact with each other. In this study, we focused on the system off-line and to provide a model used in such a system. Our model is based on the assumption that the deviations are as the outlier in the data (The deviation of the data, are minorities).

2.1 Data Structure

The target database, by a health care insurance in Iran after the identities of anonymous people, with 1,158,520 records of services provided and with Table 1 features are provided for researches. A preliminary review that we've done some of the following features is correlated with each other. Such as: patient's name and date of birth, patient name and medical centers, and so on. Since the correlation between some of the features such as the family and there was no disease, we will ignore this type of coupling features[20,21]. Considering these circumstances, the problem with the next 7 to set the sub domain becomes two-dimensional.

2.2. Methodology

Given the discussion above, the range of 7 next became a collection of two-dimensional sub domain. Incidence and risk matrices application in conjunction with the methodology flowchart is provided in Figure 1. We developed mat lab 2015a m-file based on an existing database. This application process database and create incidence matrices for all interest domains. 8 range of two-dimensional are including:

- Incidence matrix range of gender and relation
- Incidence matrix range of gender medical center
- Incidence matrix range of relation and year of birth
- Incidence matrix range of diseases and medical center
- Incidence matrix range of diseases and year of birth
- Incidence matrix range of diseases and required cost
- Incidence matrix range of the kind of cost and medical center
- Incidence matrix range of the kind of cost and required cost

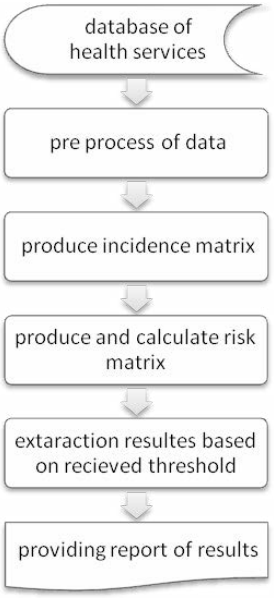

Figure 1. A schematic view of the flow chart model of the proposed algorithm.

Table 1. Describes the features in the data used

Explanation	Number of values	Type	Feature
Male and female	2	Categorical	gender
The main person insured (director) and other dependent family members, who use insurance services	9	Categorical	family relationship
Hospital and laboratory	2	Categorical	Type Price
6 different diseases that services provided to them in the database collected for investigation	6	Categorical	Disease
1.053 different medical centers that are part of the service provided	1.053	Categorical	Medical center
148 four-digit number that is representing the year of birth of the client is	148	Continuous	Year of Birth
166.942 different positive number that represents the amount of the insurance claim for the service provided	166.942	Continuous	Transfer fee

Entry (i, j) matrix of incidence is equal to the number of times that i and j trait of relevant features to meet each other in the database. About two features of the birth and the amount requested (continuous features) due to the variation trait, and the numbers are continuous, they have defined intervals. In for such a case the period is divided by five. But disadvantage is that separate parts are predefined. And may be put two serial numbers with large events in the two separate groups and cause not too see an outlier in each of the groups. So here we have used the clustering with predefined cluster; thus, by using k-means, clusters 10,20,30,50,100 tested and ultimately chose each cluster 50 to continuous features. This feature contains 148 year of birth trait became a new 50 trait.166.942 trait in the amount requested feature has also become a new 50 trait. The amount requested in the matrix name of the disease that the columns are requested amount; we have a matrix with 6 rows and 50 columns. For each case of disease i, the number of Entry (i, j) means the number of cases in which the insurance fee has been in the range of trait j. We have all matrices of incidence, and create risk matrix. The incidence matrix of a matrix sizes and in every cell, the risk assessment shows the incidence matrix equivalents at cell. With a matrix of risk, risk identification application deals with a value greater than the defined thresholds.

2.3 Risk Assessment

To assess the risk, we used incidence matrix. This matrix contains a number of events in the intersection points traits pair is optional features. As for the calculation of risk, we are compared values of different columns per row. With the relocation feature in rows and columns for each incidence matrix a risk matrix is created and each pair matrix best results are identified as high risk cases. Here the method of calculation of risk, is different based on the feature being ordered categorical or column.

2.3.1 Risk metric for Categorical Features

Gender, relation, disease, medical center, and the type of cost are un-ordered categorical features. In 16 risk matrices resulted incidence matrices In 12 cases, one of the features in the column located that they are as follows:

- The risk matrix relation-gender
- The risk matrix relation-year of birth
- The risk matrix medical center-gender
- The risk matrix medical center-disease
- The risk matrix medical center-the kind of coast
- The risk matrix gender-relation
- The risk matrix gender- medical center
- The risk matrix disease-medical center
- The risk matrix disease- year of birth
- The risk matrix disease-required cost
- The risk matrix the kind of coast- medical center
- The risk matrix the kind of coast- required cost

In the matrix, risk matrix, for example, medical center-diseases is that using the incidence matrix corresponding to the number of services provided per patient in all medical centers examined and extracts the relevant risk. Results are numbers in the range of zero and one. In this part we used the improved formula provided in. Accordingly, the maximum value in each row of incidence, at equal cell in the risk matrix, including zero and zero in the incidence matrix, replace with a risk matrix. Here outliers' numbers are close to one and numbers other than one. A disadvantage of the formula given in is that if a trait is very little in the feature column of event (For example, a family of his brother-in feature «relation», that is only 3 of the familial relationship among all the records in the database), the incidence matrix in each row and in comparison to other trait has a very small number and is identified as a high risk. Here To resolve this issue, normalization is performed on the values of each column matrix. Thus, in the formula (1) instead of comparing the number of hits in each row of incidence matrix compares the share of each trait row of feature column. Equation (1) relating to the calculation of incident matrix normalized is called MN normal incidence where the second feature (column) categorical is typed:

$$risk_M(i,j) = \frac{e^{-\left(\frac{MN(i,j)}{Max_{MN}(i)}\right)} - e^{-1}}{1 - e^{-1}} \quad (1)$$

2.3.2 Risk Metric for Ordered Features

Features also generally called continuous features have numerical values comparable with each other. Of the 7 feature used in this study, two features years are in this type year of birth and the requested cost. On 16 risks matrix has been created, in 4 cases one of the two features is in column that are as follows:

- The risk matrix year of birth- relation
- The risk matrix year of birth- disease
- The risk matrix required cost- disease
- The risk matrix required cost- the kind of cost

As we saw in the preparation stage data, amounts

each year of birth and the amount requested by clustering feature located in 50 cluster. We sort 50 cluster based on the values in them and consider cluster burst equal to 1. Due to continuous data, the center of each row determines and the distance from the center consider as well as an important factor to identify outlier. According to these points equation (2) to calculate the risk of normalized incidence matrix in the name of MN provided where the second feature (column) is of continuous type:

$$risk_{MN}(i,j) = \frac{e^{-\left(\frac{MN(i,j)}{Max_{MN}(i)} \times \left(\frac{C}{d_i(j)}\right)^3\right)} - e^{-1}}{1 - e^{-1}} \quad (2)$$

Where,

$\mathbf{V_i} = \frac{\sum_k k \times MN(i,k)}{\sum_k MN(i,k)}$ (centroid for ith rows of matrix)

and

$d_i(j) = |j - \mathbf{V_i}|$ (distance of the jth columns to the centroid of ith rows)

and

C (The number of columns of the matrix incidence)

In equation (2) in section [(C / d_i (j)) ^ 3, the amount of power 3 by trial and error, and then select different values and results were selected. However, it is possible to achieve a much better throughput, use optimization methods evolutionary algorithms.

3. Computational Results

The proposed framework code develops in mat lab 2015. Run the program and check the results several times by experts, the threshold for the number of columns and rows of the matrix that are less than 20, equal to 0.8 and in other cases, equal to 0.95 was considered as an appropriate value. The suitability of this amount is based on a tradeoff between the rate of true positive and losses in time check by the human factor. The program on a health care insurance database with 1,158,520records of services was provided. The type of system used was a Core i7 PC with a Memory 12 GB.Time of run program on all dataset, was equal to 67 seconds. By implementing this program, the cases that were much more risky than the threshold value were identified. With the implementation of the program, 81 points among all incidence risk obtained that led to the identification 4.089 dubious record. The 81 relationship that the model identified as high risk along with 150 other random relationships - that model did not identify them as high risk at a meeting of experts was held 2 expert in the field of medical and insurance. By comparing the results of the system and experts, we have seen that out of 81 points incidence, 75 cases have been properly identified as high risk (true positive) and 6 cases- the five of them about the relationship matrix between year of birth and disease, and one was the matrix of the relationship between gender and relation experts were not considered as high risk (false positive). Also among the 150 relations that model, evaluate their risk lower than the threshold only one vote from the matrix relationship between gender and relation the experts have identified as high risk. The results are presented in Table 2.

We compare our model has been implemented with the system presented in. If you do not use the clustering proposed in our model, require cost-disease and amount of cost- the kind of cost each one will has 166.942 columns; that actually would not be acceptable to us. Therefore, we performed systems based on the same cluster data. In this case, the program is closed to our model time (71 seconds) is executed and 124 points incidence identified as high risk. Most of the material that has been identified in the system

Table 2. Performance indicators

TP= 75, FP= 6, TN= 149, FN= 1		
Performance indicators	Explanation	Performance
False positive rate	$\frac{FP}{Totalnumberofinstances}$	2.60%
False negative rate	$\frac{FN}{Totalnumberofinstances}$	0.43%
True positive rate	$\frac{TP}{TP+FN}$	98.68%
Agreement rate (accuracy)	$\frac{TP+TN}{Totalnumberofinstances}$	96.97%

and in our model were not selected, were related to matrix that had one of the features «relation" or "medical center" This increase results led to a intense increase in false positive.

TP= 73, FP= 51, TN= 149, FN= 1

FPR= 18.61%, FNR= 0.36%, TPR= 98.64%, ACCURACY= 81.02%

4. Conclusion

In this research, we look to provide a framework that automatically and at high speed to detect healthcare fraud and abuse in a large data set of deals. Methodology provided by us, based on convert 7 dimensional features' domain to a collection of 2 dimensional sub-domains. This methodology involves the construction incidence matrices for each of the domains, and doing distance is based data mining. The data mining approach, risk matrices based on incidence matrix is created and the risk of each pair relationship and define with values between zero and one. By setting the threshold values can identify high is risked cases. The proposed approaches have been evaluated objectively using a real-world data set gathered from one of the healthcare insurance in Iran. This dataset with 7 feature and record 1,158,520 provided an opportunity to discuss the proposed model performance. With performance measurements with a true positive rate of 98.68%, a false positive rate of 2.60% and an accuracy of 96.97% can conclude that the proposed model performs well in this kind of problem on this type of data.

5. References

1. Aral KD, Guvenir HA, Sabuncuoglu I, Akar AR. A prescription fraud detection model. Computer Methods and Programs in Biomedicine. 2012 Apr; 106(1):37-46.
2. Shen Z, Yu H, Miao C, Weng J. Trust-based web service selection in virtual communities. Web Intelligence and Agent Systems. 2011 Aug; 9(3):227-38.
3. The Challenge of Health Care Fraud. Available from: https://www.nhcaa.org/resources/health-care-anti-fraud-resources/the-challenge-of-health-care-fraud.aspx
4. Health Care Fraud. Available from: https://www.fbi.gov/news/stories/health-care-fraud-takedown
5. Kang H, Hong J, Lee K, Kim S. The effects of the fraud and abuse enforcement program under the National Health Insurance program in Korea. 2010 Apr; 95(1):41-9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19939490
6. Shin H. A scoring model to detect abusive billing patterns in health insurance claims. Expert Systems with Applications. 2012 Jun; 39(8):7441–50.
7. Weng X, Shen J. Detecting outlier samples in multivariate time series dataset. Knowledge-based Systems. 2008 Dec; 21(8):807–12.
8. Potts DM, Addenbrooke TI. A structure's influence on tunneling-induced ground movements. Proceedings of the Institution of Civil Engineers; 1997 Apr. p. 109-25.
9. Sokol L, Garcia B, Rodriguez J, West M, Johnson K. Using data mining to find fraud in HCFA health care claims. Top Health Information Management. 2001 Aug; 22(1):1–13.
10. Precursory steps to mining HCFA health care claims. Available from: http://ieeexplore.ieee.org/document/926570/?reload=true&arnumber=926570
11. He H, Wang J, Graco W, Hawkins S. Application of neural networks to detection of medical fraud. Expert Systems with Applications. 1997 Nov; 13(4):329–36.
12. Popowich F. Using text mining and natural language processing for health care claims processing. ACM SIGKDD Explorations: Natural Language Processing and Text Mining. 2005 Jun; 7(1):59–66.
13. Yang WS, Hwang SY. A process-mining framework for the detection of healthcare fraud and abuse. Expert Systems with Applications. 2006 Jul; 31(1):56-68.
14. Virdhagriswaran S, Dakin G. Camouflaged fraud detection in domains with complex relationships. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and data Mining; USA. 2006 Aug. p. 941–7.
15. Iyengar VS, Hermiz KB, Natarajan R. Computer-aided auditing of prescription drug claims. Health Care Management Science. 2014 Sep; 17(3):203-14.
16. Thiyagarajan VS. Platfora Method for high data delivery in large datasets. Indian Journal of Science and Technology. 2015 Dec; 8(33):1-13.
17. Toward Patient Specific Long Lasting Metallic Implants for Mandibular Segmental Defects. Available from: file:///C:/Users/laptop/Downloads/My%20dissertation.pdf
18. Moghaddam NS, Elahinia M, Miller M, Dean D. Enhancement of bone implants by substituting nitinol for Titanium (Ti-6Al-4V): A modeling comparison. In American Society of Mechanical Engineers Conference on Smart Materials, Adaptive Structures and Intelligent Systems; 2014 Sep. p. 1-4.
19. Esfahani SN, Andani MT, Moghaddam NS, Mirzaeifar R, Elahinia M. Independent tuning of stiffness and toughness of additively manufactured titanium-polymer composites: Simulation, fabrication and experimental studies. Journal of Materials Processing Technology. 2016 Dec; 238:22-9.
20. Moghaddam NS, Skoracki R, Miller M, Elahinia M, Dean D. Three dimensional printing of stiffness-tuned, nitinol skeletal fixation hardware with an example of mandibular segmental defect repair. Procedia the Chartered Insolvency and Restructuring Professional. 2016; 49:45-50.
21. A numerical simulation of the effect of using porous superelastic Nitinol and stiff Titanium fixation hardware on the bone remodeling. Available from: https://spie.org/Publications/Proceedings/Paper/10.1117/12.2222075?origin_id=x4325&start_volume_number=9800