ELSEVIER

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 62 (2012) 989 – 994

Procedia
Social and Behavioral Sciences

WC-BEM 2012

A fraud detection approach with data mining in health insurance

Melih Kirlidog[a,b*], Cuneyt Asuk[b]

[a]*North-West University, Vanderbijlpark, South Africa*
[b]*Marmara University, Istanbul, Turkey*

Abstract

Fraud can be seen in all insurance types including health insurance. Fraud in health insurance is done by intentional deception or misrepresentation for gaining some shabby benefit in the form of health expenditures. Data mining tools and techniques can be used to detect fraud in large sets of insurance claim data. Based on a few cases that are known or suspected to be fraudulent, the anomaly detection technique calculates the likelihood or probability of each record to be fraudulent by analyzing the past insurance claims. The analysts can then have a closer investigation for the cases that have been marked by data mining software.

© 2012 Published by Elsevier Ltd. Selection and/or peer review under responsibility of Prof. Dr. Hüseyin Arasli
Open access under CC BY-NC-ND license.

Keywords: Data mining; health insurance; fraud detection; anomaly detection; support vector machine (SVM)

1. Introduction

One of the most important problems of the insurance industry is fraud which causes substantial losses. Gill et al. (1994) define fraud in the insurance industry as "knowingly making a fictitious claim, inflating a claim or adding extra items to a claim, or being in any way dishonest with the intention of gaining more than legitimate entitlement." Although it is difficult to estimate the amount of losses caused by fraud in Turkey, these losses are accrued into insurees as higher premiums. It is estimated that fraud in property and casualty insurance costs the Canadian insurance industry 1.3 billion Canadian Dollars every year which translates to about 10-15% of the claims paid out in Canada (Gill, 2009). It is reported that fraud detection is difficult and is not "cost effective," because if it is done incorrectly it may irritate legitimate customers and it may result in delayed claims adjudication. High costs of investigations are also a concern. As a result, many insurance companies prefer to pay the claim without investigation, because ultimately this is cheaper for them (ibid). The Association of British Insurers suggests that fraudulent claims cost the UK insurance industry over £1 billion a year and fraudsters continuously develop new types of "scams" (Morley et al., 2006).

Fraud can be seen in all insurance types including health insurance. Fraud in health insurance is realized by intentional deception or misrepresentation for gaining some shabby health benefit. Data mining tools and techniques can be used to detect fraud in large sets of insurance claim data. One of the most common data mining techniques used for finding fraudulent records is *anomaly detection*. This technique aims to detect outliers or anomalies which deviate from the usual patterns. For example, an anomaly in data traffic pattern in a computer network could be an indication of sending out sensitive data from a hacked computer (Kumar, 2005) and an anomaly in an MR image

[*] Corresponding author. Tel.: +90-216-3472859; fax: +90-216-3472859.
E-mail address: melihk@marmara.edu.tr .

1877-0428 © 2012 Published by Elsevier Ltd. Selection and/or peer review under responsibility of Prof. Dr. Hüseyin Arasli
Open access under CC BY-NC-ND license. doi:10.1016/j.sbspro.2012.09.168

could be an indication of a malignant tumor (Spence et al., 2001). Beyond the insurance industry, anomaly analyses are applied for fraud detection in diverse areas such as in credit cards, mobile phones, and insider training monitoring (Chandola et al., 2009).

Unlike most other computer software, data mining systems do not indicate the occurrence of an event with mathematical accuracy. Based on a few cases that are known or suspected to be to be fraudulent, the anomaly detection technique calculates the likelihood or probability of each record to be fraudulent by analyzing the past insurance claims. The analysts can then have a closer investigation for the cases that have been marked by data mining software.

Data mining techniques are used to detect patterns in large amount of data. Part of the larger concept of the *knowledge discovery*, data mining involves statistical, mathematical, artificial intelligence and machine learning techniques to extract and identify useful information and subsequent knowledge from large databases (Turban et al., 2005). There are several different data mining functions that use specific algorithms. Some of these functions and algorithms are presented in Table 1 along with application examples (Bigus, 1996).

Table 1. Data mining functions and algorithms

Data mining function	Algorithm	Application example
Associations	Statistics, set theory	Market basket analysis
Classification	Decision trees, neural networks	Risk assessment
Clustering	Neural networks, discriminant analysis	Market segmentation
Modeling	Linear and nonlinear regression	Sales forecasting
Sequential patterns	Statistics, set theory	Market basket analysis over time

This article will discuss the types of fraud in Turkish health insurance industry and possible data mining techniques to detect them.

2. Some common fraud types in health insurance in Turkey

An overwhelming majority of fraud events in insurance industry follow a limited number of patterns which are usually known to the insurance experts. Different types of insurance transactions can have different types of fraud. Fraud in health insurance can be specific to each country taking advantage of inadequacy of the relevant legislation or being affected by the local culture. For example, people in the countries with a collectivist culture may have a higher tendency to abuse the system compared to the countries with individualistic culture. Personal and family ties are stronger in the former compared to the latter and an uninsured person may unlawfully get insurance benefit disguising himself as an insured person. This, of course, requires the consent of the genuinely insured person.

A specific pattern that is believed have some propensity to fraud is a heuristic and based on company experience. Although each company has its own set of such patterns, those patterns usually overlap. However, the companies are usually reluctant to disclose these patterns because they are concerned for fraudsters being aware of them (Morley et al., 2006).

Insurance claims that match the known patterns can be easily detected by traditional database reporting tools or computer languages like SQL. However, this technique provides only a rough guide to insurance experts, because only a small minority of such claims is indeed fraudulent. Hence, all claims that match the known fraudulent patterns need to be closely investigated by experts. This investigation may target not only the insured individuals, but also the business partners such as insurance agencies, hospitals (health centers) or pharmacies. Sometimes the fraud may take place by the collaboration of different entities. It may even be committed by the insurance company employees.

Some known fraud types in health insurance sector in Turkey are as follows:

- Charging excessive prices for a treatment or medicine in a health center.
- Unusually high number of invoices for a particular insuree in short time frame (3-4 days).
- Insurance transaction(s) where the insuree has got some treatment or medicine but either has not paid any installments or has paid only the first installment.
- Cases where the insuree buying medicine without medical examination.
- Claiming medical invoices with dates prior to or after than the beginning of the insurance period (this is permitted in some cases).
- Excessive number of medicine claims in a specific period.
- Bank account number changes of a business partner such as agency, health center or pharmacy.
- Excessive numbers of manual invoice demands whose amounts are smaller than the usual inspection limit.
- Claims whose payable amounts are greater than the invoice amounts that insurance company will pay.

3. Data collection and research method

Database of a Turkish insurance company was used in this research. The database contained detailed claim records as well as other necessary information such as business partners and customers. Anomaly detection analysis was performed on an Oracle system that uses *support vector machine (SVM)* algorithm. SVM is basically a classification technique that works in a one-class setting where individual records are identified as normal or anomalous (Vapnik, 1995). The system is “trained” to determine a boundary between normal and anomalous records. Then each record is compared with that boundary and is identified either as normal or anomalous. SVM is a kernel-based algorithm where kernel transforms the input data to a high-dimensional space to solve the problem. Oracle 11g Release 1 which was used in this research uses *Gaussian (nonlinear)* or *Linear* kernels in data mining process. The linear kernel function reduces the cases to a linear equation on the original attributes in the training data whereas Gaussian kernel transforms the cases to individual points in the n-dimentional space on which it attempts to separate the points into subsets with homogeneous target values. Although the Gaussian kernel uses nonlinear separators, it constructs a linear equation within the kernel space. Linear kernel was used in this research.

3.1. Data preparation

The claims were recorded in two relational tables in the database. One of them contained the claim header and the other contained detailed claim records. The mining activity was mainly performed on the claim header file because it had a richer source of information compared to the claim detail file which contained only money-related records such as payable amount, paid amount, net amount, and invoice amount. Where necessary, data from the claim detail table were also used.

There were 808348 records in the claim header table covering a time range from 2001 to 2009. Although a small minority of the records in that table was known to be erroneous with zero of null fields, none of them were deleted before the data mining process. The reason for that was the fact that data mining software automatically handled such records.

4. Findings and discussion

The data mining software calculates the probability of the anomaly of each record. If the probability is greater than 50% then the record is marked as anomalous. The software identified 6595 claim header records with probabilities ranging from 50.0% to 67.3%. The anomalous records were analyzed according to several criteria three of which are explained in this article.

4.1. Criteria one: Rejected claims

There were 480 rejected claim records in the table. A claim may be rejected for several reasons including suspected fraud. 147 of the rejected records were also marked by data mining software as anomalous. Given the very high number of claim records, this concurrence where 147 records are identified as problematic both by humans and the computer is striking and it can be regarded as a justification of 30.6% of the rejected claims.

Table 2. Anomalous and rejected claims

Total claim records	Anomalous claim records	% Anomalous claim records	Rejected claim records	% Rejected claim records	Probability of a rejected claim to be anomalous
808348	6595	0.816%	480	0.059%	30.6%

4.2. Criteria two: Excessive claims in health center types

As stated above, an important predictor of insurance fraud in Turkey is to charge excessive amounts by a health center. Since it is difficult to do it continuously without being detected, this type of fraud can take several forms such as dividing the claim into smaller amounts or doing it in from time to time with long time intervals between frauds.

The insurance company works with seven types of health centers. Table 3 presents these types along with other data such as average amount of claims paid to them from 2001 to 2009. The first line displays the number of health centers the company has agreement with. The next line shows the number of claims for all health center types. Total number of claims is slightly different than the record number in the claim header table due to the difference between claim header and claim detail tables.

The next three rows show the average of total, non-anomalous and anomalous records, respectively. The differences between total and non-anomalous averages are negligible for all health center types due to the reason that over 99% of the records are non-anomalous. However, there are important differences between anomalous and non-anomalous records in all types. Average amounts in hospitals, pharmacies, freelance doctors and polyclinics are higher for anomalous records. Investigation of those records can provide important clues for detecting fraudulent claims.

The sixth row shows the highest claim amount for all health center types. Those high numbers are normally closely scrutinized by the insurance experts. The next line shows the number of anomalous records in the highest one hundred claims in each type. These records are particularly prone to fraud due to two factors. Firstly, they are high in amount like most fraudulent cases and secondly their anomalous nature brings attention to them. Close inspection of those records can also bring some more clues. For example, the three anomalous records in the “others” type are 2585 TL each and are from the same insuree who claimed them from a particular health center in the same day. It is probable that the possibly fraudulent claim has been divided into three installments in order to avoid the attention due to high amount.

Table 3. Health center types, average claim amounts in TL and claim numbers

	Hospitals	Pharmacies	Freelance doctors	Polyclinics	Diagnostic centers	Medical visualization centers	Others
# health centers	743	249	8491	3747	377	558	5265
# claims	331200	296076	37499	117371	13007	13206	4050
General average of claims (TL)	376.7	40.6	218.1	184.2	182.9	337.2	417.0
Non-anomalous claims average (TL)	376.3	40.6	217.5	184.1	183.6	337.7	418.5
Anomalous claims average (TL)	445.0	62.5	259.1	188.8	137.5	231.7	347.9
Highest claim amount (TL)	224791.2	10277.1	14784.1	183048.9	9260.7	13314.1	27830.6
# anomalous in highest 100 claims	0	0	4	0	2	0	3

4.3. Criteria three: Excessive claims in health centers

The last investigation was the analysis of high amount of claims by individual health centers. To this end, claims over 2000 TL from the anomalous records were drawn from the database. A total of 120 records belonging to 55 health centers were identified in this way. The target was further narrowed by taking health centers that have more than three anomalous records. Table 4 shows the ten health centers identified, number of anomalous claims and total amounts for those claims (excluding claims equal or less than 2000 TL) as a result of this analysis. Some further and detailed analysis of these claim records could reveal dubious claims made by these health centers.

Table 4. Health centers which have more than three anomalous records for the claims exceeding 2000 TL

	# anomalous claims > 2000 TL	Total amount for these claims (TL)
Health center 1	13	42744.09
Health center 2	4	30045.28
Health center 3	4	27633.61
Health center 4	6	30898.81
Health center 5	6	40829.09
Health center 6	13	39277.32
Health center 7	7	43787.79
Health center 8	5	13498.02
Health center 9	5	32185.25
Health center 10	4	25401.90

5. Conclusion

Data mining methods such as anomaly detection, clustering and classification can successfully detect anomalies or outliers in large sets of data. This can be very useful for insurance industry which has problems with fraudulent claims. Once the anomalous claims are detected, several analyses must be made on them in order to conduct a thorough investigation. The main task in these analyses are to narrow the target for detecting frauds. Although most fraud patterns are usually known by insurance experts, such an investigation can also reveal some new and unknown patterns.

Usually insurance companies have claim data for large time frames. Although this article used longitudinal data for nine years, analyses for shorter time frames such as one year must also be made. These shorter analyses can be useful for “hit-and run” type of frauds which can be difficult to detect in long time frames.

Fraud detection and prevention can be beneficial for consumers who have to pay to the fraudsters in the form of higher insurance premiums. Technology for this task is available today and the insurance experts need to be trained for using it effectively.

References

Bigus, J. P. (1996). *Data Mining with neural networks*. New York: McGraw-Hill.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 15/1-15/58.

Gill, K. M., Woolley, K. A., & Gill, M. (1994). Insurance fraud: The business as a victim. In M. Gill (Ed.), *Crime at work, Vol 1. (pp. 73-82).* Leicester: Perpetuity Press.

Gill, W. (2009). Fighting fraud with advanced analytics. *Canadian Underwriter*, September, 28-32.

Kumar, V. (2005). Parallel and distributed computing for cybersecurity. *IEEE Distributed Systems Online,* 6(10), 1-9.

Morley, N. J., Ball, L. J., & Ormerod, T. C. (2006). How the detection of insurance fraud succeeds and fails. *Psychology, Crime & Law*, 12(2), 163-180.

Spence, C., Parra, L., & Sajda, P. (2001). Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model. *In Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*. IEEE Computer Society, 3-10.

Turban, E., Aronson, J. E., & Liang, T. (2005). *Decision support systems and intelligent systems*. (7th ed.). Upper Saddle River, NJ: Pearson Education International, (Chapter 5).

Vapnik, V. N. (1995). *The nature of statistical learning theory.* New York: Springer-Verlag.