A Purview of the Impact of Supervised Learning Methodologies on Health Insurance Fraud Detection

Ananthi Sheshasaayee and Surya Susan Thomas(✉)

Department of Computer Science, Quaid-e-Millath Government College for Women, Chennai, Tamil Nadu, India
{ananthi.research, susann.research}@gmail.com

Abstract. A plethora of researches is happening in almost all sectors of insurance to improve the vitality and vibrance of its existence. As years pass, the volume of insurance policy holders increases which is directly proportional to the occurrence of frauds in these sectors. The presence of fraud is always an obstacle to the growth of an insurance organization. This paper confers the various supervised learning methodologies employed in detecting health insurance frauds.

Keywords: Health insurance · Fraud detection · Data mining · Supervised learning

1 Introduction

The health insurance sector has taken a high rise in the recent years due to the impact of the vulnerability to get hospitalized and also of the mounting hospital expense the patient had to concede. This paradigm has led nearly everyone to take health insurance policies. It gained popularity in the developing countries like India and China in the late years. This sector too faced hindrance like fraud, which retarded the profit of insurance companies, and this then paved the way for researchers to identify strategies to halt and mitigate frauds at the earliest.

Every year, millions of dollars are depleted from the insurance providers due to frauds. In order to sustain the profit, the insurance companies raise the premium amount and this in turn affects the genuine policy holders too [1]. It is estimated that frauds steal up to fifteen percent of the taxpayer amount that is used to fund the government-aided health care, making it crucial for the government agencies to find some cost-effective methods to pinpoint the fraud claims and transactions.

It is tough to eliminate these frauds and fraudsters completely, but it can be easily detectable with the incorporation of data mining methods and techniques along with artificial intelligence [2]. This mining of data and drawing patterns to recognize the odd ones gained popularity from the early 2000s.

© Springer Nature Singapore Pte Ltd. 2018
V. Bhateja et al. (eds.), *Information Systems Design and Intelligent Applications*, Advances in Intelligent Systems and Computing 672,
https://doi.org/10.1007/978-981-10-7512-4_98

1.1 Classification of Health Insurance Frauds

Fraud in healthcare industry is just like in any other industry [3]. Fraudsters obtain full benefit of unjust profit with the help of healthcare crooks which includes patients, payers, vendors, suppliers, employers, and healthcare providers including pharmacists (Table 1).

Table 1. Illustration of various types of healthcare insurance frauds [4]

S No	Type of fraud	Area
1	Billing services not rendered	Hospital/clinic
2	Billing uncovered service as utilized service	Hospital/clinic
3	Altering dates of assistance	Hospital/clinic
4	Altering location of assistance	Hospital
5	Altering provider of assistance	Patient/customer
6	Incorrect reporting of analysis (unbundling)	Hospital/patient
7	Overutilization of assistance	Hospital/patient
8	Kickbacks/Bribery	Hospital/doctors
9	False/unnecessary issuance of drugs	Pharmacists
10	Up coding or down coding	Hospital/clinic

1.2 Statistical Insight into the Impact of Healthcare Frauds

Figure 1 shows criminal healthcare prosecutions over the last 20 years in the USA according to the TRAC report [5].

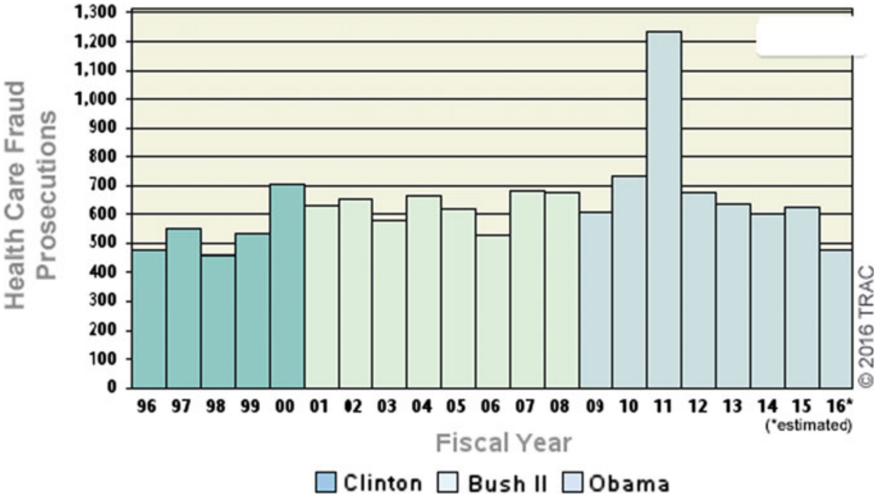

Fig. 1. Statistical representation of criminal healthcare prosecutions over the last 20 years in the USA

2 Learning Methodologies to Detect Health Insurance Frauds

Exploring data or simply “data mining” is a constant progress which involves discovery of a new fact or a hidden truth either through automated or manual procedures. It is a powerful mechanism in an exploratory research scenario where there are no fixed beliefs about the future outcomes. Data mining is the quest for new, variant, and nontrivial truths from large quantity of data. It is literally a combined effort of data analysts and the computers [6]. Solutions are obtained by harmonizing the knowledge of data analysts and the feedbacks of the search effectiveness of the computers.

2.1 Data Mining

Data mining has become extensively popular or to be precise very essential in the healthcare domain. Its application has become beneficial for all the factions in the healthcare domain; for instance, it can help health insurance providers to find hoaxs, healthcare enterprises to offer better customer relationship relations, doctors to recognize better treatment for their clients, and in course patients have more affordable services. A copious amount of data is generated here, and the traditional methods to handle this data are complicated and arduous [7]. Data mining provides efficient methods and techniques to convert these immense data into useful information for decision making.

2.2 Healthcare Data

Data used in health care can be classified mainly into four groups: [8]

(a) Hospital data (health record data of patients, medical images, laboratory and surgery reports, etc.)
(b) Patient conduct data (data collected through monitors and other wearable devices)
(c) Pharmaceutical data (medicines provided to the patients, etc.)
(d) Health insurance claim data (data includes services provided to the patient, their payment details, etc.)

Figure 2 shows the types of healthcare data.

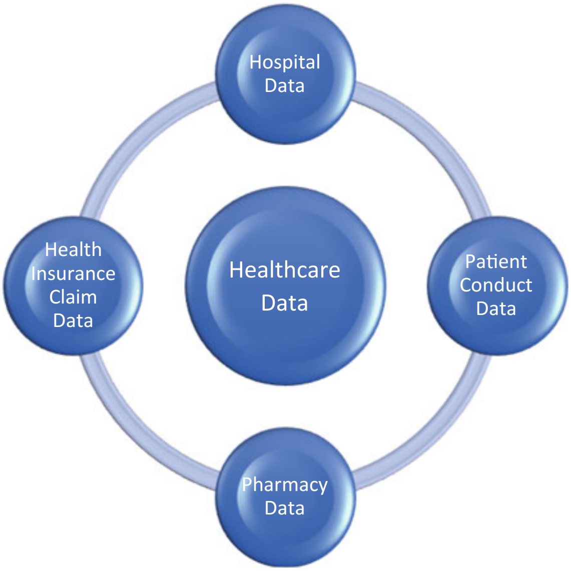

Fig. 2. Different types of healthcare data

2.3 Supervised Learning and Unsupervised Learning

The two preeminent learning methodologies are supervised learning and unsupervised learning methodologies, with the supervised learning using the trained data sets to perform mining of data while the unsupervised learning using the raw data or the real data. Fraud detection using supervised data sets is found to be more efficient and accurate [9]. But the adversity in obtaining trained data has led researchers to use raw data, i.e., using unsupervised learning. The outcome was found to be less competent. So, it is observed that supervised learning methodology has set its cardinality in detecting frauds with more effectiveness.

3 Supervised Learning in Health Insurance Fraud Detection

Supervised learning is a prominent data mining technique which has a reliable variable that is utilized either for a classification or prediction from a group of self-reliant attributes. Samples of supervised learning methodologies are naïve Bayes, linear regression, decision trees, random forest, logistic regression, and support vector machines [10]. The impediment in this type of research is that it needs some predictors or class labels to mold data for classification or prediction.

Health insurance claim studies are mostly done using supervised learning methodology (SLM). Figure 3 shows the claim process flow using SLM. Insurance claims are recorded in the database and are processed with the help of supervised learners, and claims are subjected to fraud detection models, where fraudulent claims are triggered while reports for genuine claims are generated and are subjected to payment clearing. The fraudulent claims are validating again and given to the fraud-mitigating team for further processing.

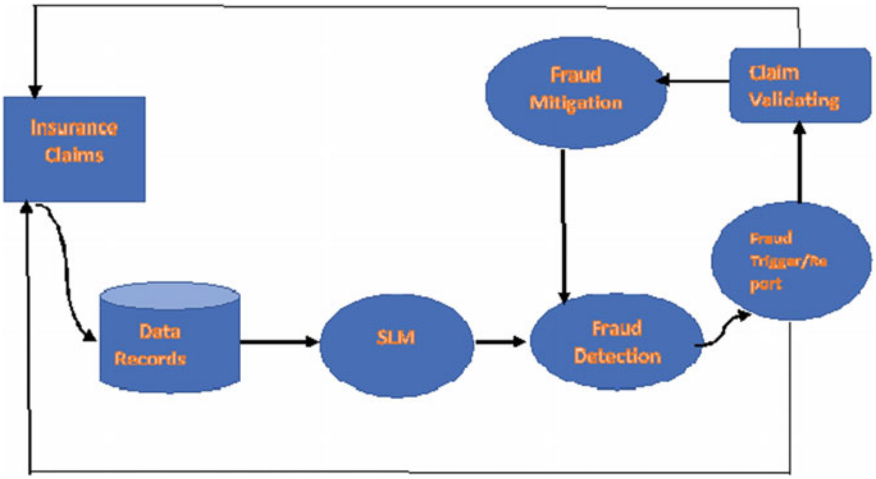

Fig. 3. Claim process flow using SLM

4 Related Works

Many studies have been carried out, implementing supervised learners to detect insurance claim frauds. This paper brushes through the works which throws an insight into the fraud detection mechanisms optimized in the insurance sector to catch hold of the fraudulent claims.

Travaille et al. [11] examine “Medicaid” fraud detection by evaluating techniques implemented in various sections from telecommunications to credit cards to healthcare that uses many machine learning techniques. They also asserted that since the divergence of health insurance data is plentiful, trained information is not easily acquired when compared to the availability in other fields.

Liu et al. [12] handle supervised learners such as multilayer perceptron (MLP) and decision trees for fraud detection in healthcare and lays a double-fold frame of anomaly revelation ideas using geolocation. In addition to that, the authors draft a supervised learning process using genetic algorithms and k-nearest neighbor (kNN) algorithm. The authors incurred that supervised learning yields more efficiency but has difficulty in procuring labeled data.

Phua et al. [13] had an in-depth study on the computerized deceit detection works collected from a span of a decade. The authors analysed that several research studies focuses mainly on complicated, nonlinear supervised algorithm whereas lighter methods such as naïve Bayes produce better and in some cases more efficient results.

Johnson et al. [14] framed an anomaly detection by implementing a multistage access to highlight hospital care fraud apparently using both private and public available data. A risk ranking is then formulated with the help of decision trees.

Joudaki et al. [15] guide an examination on the mining data learning processes fixating mainly on healthcare fraud detection. Their work discusses both supervised and unsupervised learning algorithms in the fraud detection field and mentions the need to refine supervised models to better efficiency.

Kumar et al. [16] add to the research work in fraud detection by framing a support vector machine (SVM) supervised learning for prediction of errors in insurance claims.

Ngufor et al. [17] converge to the investigation of provider fraud, concentrating in obstetrics claims, using unsupervised learners such as outlier detection with supervised learners like regression classification.

Brockett et al. [18] confined to the study considering the number of provider visits, their next visits, and the scans done without inpatient costs as the model criteria. They examined each criterion using a "ridit" scoring. Then, they utilized a double-way classification that separated claims into genuine and fraudulent.

Ortega et al. [19] suggested a many layered feed-forward neural networks to identify fraudulent claims. They used a false trigger cost as the model variable.

Liou et al. [20] proposed a model which dealt with nine types of cost-related parameters in their mining data process. They injected a classification tree, logistic regression, and neural network algorithm to spot hoax in diabetic disease claims. It was observed that the introduction of classification tree produced better results.

5 Conclusion

A wide spectrum of researches are carried out over the years to help expose frauds in the insurance field. Analysts are in search of newer and better methods to detect frauds and thus help the insurance companies to combat fraudsters in an efficient manner. Supervised and unsupervised learning algorithms are used in many cases, where both have their own advantages and disadvantages. Supervised learning has more accuracy points, but the effort to obtain labeled data is onerous. On the other hand, unsupervised learning has some negative aspect of an uncertainty since the main connection between the measured attributes of the information is unexplored. Hence, it is observed that a hybrid learning mechanism is recommended in dealing with hoax detection in the health insurance domain.

Acknowledgements. My sincere gratitude to my research guide Dr. Ananthi Sheshasayee for her expert guidance and valuable suggestions for the formulation of this paper. I thank all my well-wishers for their support and help.

References

1. Copeland, Leanndra, et al. "Applying business intelligence concepts to medicaid claim fraud detection." *Journal of Information Systems Applied Research* 5.1 (2012): 51.
2. Cortesão, Luis, et al. "Fraud management systems in telecommunications: a practical approach." *Proceeding of ICT*. 2005.
3. Berwick, Donald M., and Andrew D. Hackbarth. "Eliminating waste in US health care." *Jama* 307.14 (2012): 1513–1516.
4. http://www.fraud-magazine.com/article.aspx?id=4294976280.
5. http://trac.syr.edu/tracreports/crim/424/.

6. Gnanapriya, S., et al. "Data Mining Concepts and Techniques." *Data Mining and Knowledge Engineering* 2.9 (2010): 256–263.
7. Koh, Hian Chye, and Gerald Tan. "Data mining applications in healthcare." *Journal of healthcare information management* 19.2 (2011): 65.
8. Chandola, Varun, Sreenivas R. Sukumar, and Jack C. Schryver. "Knowledge discovery from massive healthcare claims data." *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013.
9. Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.
10. Bauder, Richard, Taghi M. Khoshgoftaar, and Naeem Seliya. "A survey on the state of healthcare upcoding fraud analysis and detection." *Health Services and Outcomes Research Methodology*: 1–25.
11. Travaille, Peter, et al. "Electronic fraud detection in the US medicaid healthcare program: lessons learned from other industries." (2011).
12. Liu, Qi, and Miklos Vasarhelyi. "Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information." 29th World Continuous Auditing and Reporting Symposium (29WCARS), Brisbane, Australia. 2013.
13. Phua, Clifton, et al. "A comprehensive survey of data mining-based fraud detection research." *arXiv preprint* arXiv:1009.6119 (2010).
14. Johnson, Marina Evrim, and Nagen Nagarur. "Multi-stage methodology to detect health insurance claim fraud." *Health care management science* 19.3 (2016): 249–260.
15. Joudaki, Hossein, et al. "Using data mining to detect health care fraud and abuse: a review of literature." *Global journal of health science* 7.1 (2014): 194.
16. Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." *ACM computing surveys (CSUR)* 41.3 (2009): 15.
17. Wojtusiak, Janusz, et al. "Rule-based prediction of medical claims' payments: A method and initial application to medicaid data." *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*. Vol. 2. IEEE, 2011.
18. Brockett, Patrick L., et al. "Fraud classification using principal component analysis of RIDITs." *Journal of Risk and Insurance* 69.3 (2002): 341–371.
19. Ortega, Pedro A., Cristián J. Figueroa, and Gonzalo A. Ruz. "A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile." *DMIN* 6 (2006): 26–29.
20. Liou, Fen-May, Ying-Chan Tang, and Jean-Yi Chen. "Detecting hospital fraud and claim abuse through diabetic outpatient services." *Health care management science* 11.4 (2008): 353–358.