DALHOUSIE UNIVERSITY

Faculty of Computer Science

A Survey Report on Data Mining Approaches in Healthcare Fraud Detection

CSCI 6405 – Data Mining & Data Warehousing

Name: Arka Ghosh
Banner ID: B00911033

Assignment: 01

Introduction

Data mining refers to the process of predicting the outcome by searching for anomalies, patterns and correlation within large amount of datasets [1]. In recent times, many industries have been using data mining techniques to improve their customer experience and increase their product safety, and thus the healthcare industry is no exception. Within this industry, healthcare insurance has been drawing the attention of the fraudsters since a long time. Traditional approaches cannot handle and interpret the massive volumes of data created by healthcare transactions because the data are too complicated and voluminous most of the times. Data mining offers the process and technology for turning this enormous amount of medical data into useful information for decision-making [2]. In regard to this industry, this paper surveys the development and implementation of data mining and related techniques in detecting fraud healthcare and medical insurance claims. A summary of the proposed solutions and its performance to this discipline for the selected papers has also been presented in the paper.

Problem Description

Healthcare sector has become a lucrative target for the fraudsters because of the large amount of money being invested in this sector by government and the private sector. Healthcare insurance fraud is defined as an intentional attempt of giving falsified or misleading information to an insurance company by an individual or a policyholder, knowing that that the misleading information may lead to an unauthorized advantage to them or others [3]. Anomaly in medical insurance can be divided into some common categories such as duplicating claims, fraudulent use of the medical insurance cards of other people, billing for services that are not rendered, frequent or over-utilization of a medical facility and acquiring fake invoices with the help of the physicians [4]. In 2016, The Justice Department of USA prosecuted the biggest anomaly case in the medical sector in its history by bringing charges against 301 people including physicians, nurses and other medical professionals for falsifying medical billing worth around $900 million [5]. According to Blue Cross, a health insurance coverage provider in Canada, health-care frauds accounts between two to ten percent of overall health-care expenditure in Canada, ranging from $440 million to $2.2 billion each year [6]. The statistics suggests that insurance fraud has ripped a hole in the healthcare industry. Previously, health insurance firms used to apply heuristic rules to detect the insurance frauds which were compiled from prior fraud cases and are used to identify the fraud in the healthcare insurance either by manual human inspection or by a third party. In recent times, the medical database has grew significantly in size which made this traditional method of detecting fraud vulnerable. Additionally, SQL based operations fail to identify various emerging fraud scheme in these databases. Such situation requires a semi-automated approach that can be useful in discovering hidden information in such kind of databases. Thus, data mining approaches and machine learning algorithms are vastly used in the healthcare industry to tackle the problem of healthcare insurance frauds in recent times due to its inherent characteristics of finding various known and/or hidden patterns through filtering immense amount of datasets.

Proposed Solutions

Vipula et al proposed a novel hybrid model, combing both supervised and unsupervised technique to detect fraud in the health insurance [7]. In context of medical insurance claims, supervised learning is not able to classify the claim of new diseases as it is trained using the pre-defined class labels. On the other hand, it is possible to discover both old and new type of frauds through unsupervised learning as it's not limited to any predefined class patterns but there will be missing set of features for training in such kind of learning method. Due to drawbacks of each of the learning methods, this hybrid model is proposed. This proposed model organizes medical insurance claims depending on the kind of diagnosis, and then categorizes them to locate claims that are duplicated. For this hybrid model, Evolving Clustering (ECM) method is used as the data used is dynamic and it keeps changing with respect to time. ECM clusters the incoming data points when they arrive by adjusting the cluster's size and position. After clustering the data through ECM, the data is forwarded to Support Vector Machine (SVM) which determines which class the insurance claims fall into.

Herland et al has provided a fraud label mapping approach in showing best learners to detect Medicare fraud claims in [8]. This research is conducted using three CMS datasets (Part B, Part D and DMEPOS) of 2013 to 2015 and fourth dataset named Combined Dataset which is created from the three primary CMS datasets. List of Excluded Individuals and Entities (LEIE) is used to generate fraud labels for all four datasets in this research which features physicians who have actually committed fraud in the real world, allowing for a reliable evaluation of fraud detection performance. Data imputation, deciding which variables to keep and transferring the data through aggregation to match the LEIE dataset's level in case of fraud mapping and creating the fourth dataset, is the novel data processing step used in this explanatory study. To run and validate the models used in this research, Apache Spark on top of a Hadoop YARN cluster is used due to the large size of the datasets. The datasets are trained evaluated through 5 fold cross-validation with repeating the process for ten times. The three classification models available in the Apache Spark Machine Learning Library: Random Forest (RF), Gradient Boosted Trees (GBT) and Logistic Regression (LR), and Area Under Curve (AUC) metric is used to gauge detection performance. Additionally, statistical significance are also estimated with the Analysis of Variance (ANOVA) and Tukey's Honest Significant Difference (HSD) [9] tests to add robustness around the result.

A novel approach using the concepts of sequence mining and sequence prediction to detect healthcare insurance frauds has been proposed in [10]. The proposed methodology is evaluated using the insurance claim data of a private hospital. This proposed methodology is based on two modules which are Sequence Rule Engine (SRE) and Prediction Based Engine (PBE). Sequence Rule Engine is initially used to transform the transactional data into time series sequence traces. As it is not possible to detect anomalies directly from the traces, sequence of services availed in each specialty are captured separately modifying the minimum support value based on the amount of the transactions completed in that specialty. In this methodology, Prefixspan Algorithm and Bayes Theorem is also applied to populate frequent and rare sequences in the SRE for specific specialty. Sequence matching of patient sequences is done against Sequence Rule Engine's common sequences as well as sequence rule engine's rare sequences for the detection of fraud instances. The sequences that are not complaint with the sequence rules are forwarded to PBE that treats test cases as testing data which is a vector of test sets. If a value of a test case is predicted null by the PBE, this means the particular test case is proof of fraudulent activities.

Performance of the Proposed Solutions

In the proposed model in [7], the claims filed by the patient are submitted to the Hybrid Model where ECM clusters the data which is followed by SVM that is used in detecting the fraudulent claims. As mentioned above, the data points are clustered through ECM in proposed model in [7]. There is a parameter named radius associated with the cluster which is initially set to zero. This parameter is used in determining the boundaries of the associated cluster. This parameter keeps increasing when there are more data points added to the cluster. There is another parameter called Distance Parameter that is used in determining the cluster addition. In the training phase of the SVM, there are two defined classes known as "legitimate" and "fraudulent". It selects support vectors and calculates the greatest marginal hyperplane that divides the claims into two class. After this phase, the SVM classifies into which class the incoming claims will fall into.

The proposed approach in [8] has provided highest performance for the Combined Data with LR learner model, resulting into AUC of 0.816. The next best result is the Part B dataset with LR with 0.805, while Part D and DMEPOS datasets produces the lowest AUC score. Although the Combined Dataset produces the highest overall AUC score, the Part B dataset indicates the lowest variances across the learners in detecting the fraud, with AUC score of 0.79569 with GBT and 0.79604 with RF learners. It is predicted in this paper that the positive results obtained with each dataset using LR is because of the Squared Error loss function combined with L2 regularization which penalizes big coefficients and improves generalization

performance, making LR resilient to noise and overfitting. Additionally, Tukey's HSD test on the results support the LR Learner's and Combined Dataset's positive performance in detecting Medical Fraud with no statistical differences in compared to other dataset's results.

The proposed methodology in [10] has the ability to identify the anomalous behavior from the transaction data by performing two cascaded checks at two levels. The first one is similarity verification from the Sequence Rule Engine and the later one is fraud detection using the Prediction Based Engine. The sub-set of sequence database for each medical specialty which has been transformed from the transactional data is used as input in the Sequence Rule Engine. The goal of creating the frequent sequence based rule engine is to find any anomalies in cases when a patient has had one sequence of services from a specialist for many or only one occasion. However, there's a chance that the identified anomaly isn't one at all. Taking this into account, the prediction-based engine is also created that detects fraud by predicting a group of services for each test set under consideration. This proposed methodology is able to detect four types of fraud cases namely Services Available but Not Compliant with Sequence Generate by SRE, Repetition in Availed Services, Few Services in a Sequences that are Anomalous and Services Availed from Specialty that are not permitted from this Specialty [10]. This proposed methodology has been able to provide an average accuracy of 85% in detecting fraud in each specialty.

Comments and Discussions

Healthcare frauds continue to be a great threat not only for a particular country but also for the entire globe. The three novel approaches outlined in the above mentioned research studies indicate impressive performance in tackling the problem. Although considering the advantages of using both supervised and unsupervised learning altogether, the proposed system in [7] is efficient in predicting the duplicate claims and outliners but other kinds of frauds cannot be predicted through the proposed model. The exploratory work outlined in [8] shows the best overall Medicare fraud detection performance using the Combined Dataset with LR but the problem of class imbalance, rarity, and any techniques to mitigate the adverse impacts on model performance is given less attention in this research. The methodology proposed using the concepts of sequence mining and sequence prediction in [10] indicates a very novel approach in fraud detection from the practical data-driven and performance-oriented perspective. Moreover, this methodology is able to cover more relevant fraud types.

References

[1] "What is data mining?", *Sas.com*, 2022. [Online]. Available: https://www.sas.com/en_ca/insights/analytics/data-mining.html. [Accessed: 30-Jan-2022].

[2] HC. Koh and G. Tan, "Data mining applications in healthcare." *Journal of healthcare information management: JHIM*, *19*(2), 64–72, 2015.

[3] "Fraud, Waste & Abuse FAQs", *Magellanprovider.com*, 2022. [Online]. Available: https://www.magellanprovider.com/education/fraud-waste-and-abuse/fraud-waste-abuse-faqs.aspx. [Accessed: 01-Feb-2022].

[4] J. Gill, "Health insurance fraud detection", *ERA*, 2020. [Online]. Available: https://doi.org/10.7939/r3-wvjc-sd55.

[5] M. Vasilogambros, "Justice Department Conducts Largest Healthcare Fraud Takedown in U.S. History", *The Atlantic*, 2022. [Online]. Available: https://www.theatlantic.com/news/archive/2016/06/healthcare-fraud-takedown/488306/. [Accessed: 01-Feb-2022].

[6] *Ab.bluecross.ca*, 2022. [Online]. Available: https://www.ab.bluecross.ca/pdfs/81984-fraud-member-letter.pdf. [Accessed: 01- Feb- 2022].

[7] V. Rawte and G. Anuradha, "Fraud detection in health insurance using data mining techniques," 2015 International Conference on Communication, Information & Computing Technology (ICCICT), 2015, pp. 1-5, doi: 10.1109/ICCICT.2015.7045689.

[8] M. Herland, T. Khoshgoftaar and R. Bauder, "Big Data fraud detection using multiple medicare data sources", *Journal of Big Data*, vol. 5, no. 1, 2018. Available: 10.1186/s40537-018-0138-3.

[9] J. Tukey, "Comparing Individual Means in the Analysis of Variance", *Biometrics*, vol. 5, no. 2, p. 99, 1949. Available: http://doi.org/10.2307/3001913.

[10] I. Matloob, S. A. Khan and H. U. Rahman, "Sequence Mining and Prediction-Based Healthcare Fraud Detection Methodology," in IEEE Access, vol. 8, pp. 143256-143273, 2020, doi: 10.1109/ACCESS.2020.3013962.