International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 01 | Jan 2019 www.irjet.net p-ISSN: 2395-0072

A Survey Paper on Fraud Detection and Frequent Pattern Matching in Insurance claims using Data Mining Techniques

Pinak Patel[1], Siddharth Mal[2], Yash Mhaske[3]

[1,2,3]*Asst. Professor, Geeta Sorate, Dept. of Computer Engineering, MAEER's MIT COE, Maharashtra, India*

Abstract - *Fraudulent insurance claims increase the burden on society. Frauds in health care systems have not only led to additional expenses but also degrade the quality and care which should be provided to patients. Insurance fraud detection is quite subjective in nature and is fettered with societal need. This empirical study aims to identify and gauge the frauds in health insurance data. The contribution of this insurance claim fraud detection experimental study untangle the fraud identification frequent patterns underlying in the insurance claim data using rule based pattern mining. This experiment is an effort to assess the fraudulent patterns in the data on the basis of two criteria period based claim anomalies and disease based anomalies. Rule based mining results according to both criteria are analysed. Statistical Decision rules and k-means clustering are applied on Period based claim anomalies outliers detection and association rule based mining with Gaussian distribution is applied on disease based anomalies outlier detection. These outliers depict fraud insurance claims in the data. The proposed approach has been evaluated on real-world dataset of a health insurance organization and results show that our proposed approach is efficient in detecting fraud insurance claim using rule based mining.*

***Key Words*: Data mining; Health Insurance claim, Rule Based Mining, Frequent patterns.**

1. INTRODUCTION

Health care has become a major expenditure in most of the countries. The large amount of money involved in this sector had made it as a target for frauds. According to the National Health Care Anti-Fraud Association, health care fraud is an intentional deception or misrepresentation made by a person, or an entity that could result in some unauthorized benefit to him or his accomplices [1]. Health care abuse is produced when either the provider practices are inconsistent with sound fiscal, business or medical practices, and result in an unnecessary cost or in reimbursement of services that are not medically necessary or that fail to meet professionally recognized standards for health care.

Moreover, effective fraud detection techniques and models are needed to improve the quality and reducing the cost of health care services, for which expertise domain knowledge is required. In recent years, various systems have been implemented to identify different types of fraud but services provided are limited due to small set of predefined rules specified by domain experts.

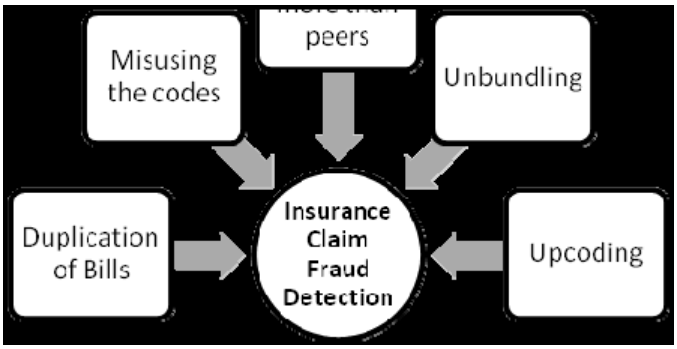

Fig. 1: Generalized Fraud Detection Model

Researches worked on various domains to detect frauds such as credit card fraud, money laundering, telecommunication fraud, computer intrusion and scientific fraud using data mining techniques. According to recent advancement in insurance claim policies, fraud detection system requires more sophisticated methods those are capable to automatically learn the fraud pattern from data using data mining (DM) [19], machine learning (ML) or statistical model. On the other end, paucity of usage of these techniques in this direction have drawn our attention to accord such research problem. Advanced and novel insurance claim policies transformed it to complex and challenging research problem. This research work introduces insurance claim fraud detection approach to identify frauds in insurance claim policy.

2. LITERATURE SURVEY

2.1 Big Data Fraud Detection using multiple medicare data sources

In the United States, advances in technology and medical sciences continue to improve the general well-being of the population. With this continued progress, programs such as Medicare are needed to help manage the high costs associated with quality healthcare. Unfortunately, there are individuals who commit fraud for nefarious reasons and personal gain, limiting Medicare's ability to effectively provide for the healthcare needs of the elderly and other qualifying people. To minimize fraudulent activities, the Centers for Medicare and Medicaid Services (CMS) released a number of "Big Data" datasets for different parts of the

Medicare program. In this paper, we focus on the detection of Medicare fraud using the following CMS datasets: (1)

Medicare Provider Utilization and Payment Data: Physician and Other Supplier (Part B), (2) Medicare Provider Utilization and Payment Data: Part D Prescriber (Part D), and (3) Medicare Provider Utilization and Payment Data: Referring Durable Medical Equipment, Prosthetics, Orthotics and Supplies (DMEPOS). Additionally, we create a fourth dataset which is a combination of the three primary datasets. We discuss data processing for all four datasets and the mapping of real-world provider fraud labels using the List of Excluded Individuals and Entities (LEIE) from the Office of the Inspector General.

Our exploratory analysis on Medicare fraud detection involves building and assessing three learners on each dataset. Based on the Area under the Receiver Operating Characteristic (ROC) Curve performance metric, our results show that the combined dataset with the Logistic Regression (LR) learner yielded the best overall score at 0.816, closely followed by the Part B dataset with LR at 0.805. Overall, the Combined and Part B datasets produced the best fraud detection performance with no statistical difference between these datasets, over all the learners. Therefore, based on our results and the assumption that there is no way to know within which part of Medicare a physician will commit fraud, we suggest using the Combined dataset for detecting fraudulent behaviour when a physician has submitted payments through any or all Medicare parts evaluated in our study.

Datasets

In this section, we describe the CMS datasets we use (Part B, Part D and, DMEPOS). Furthermore, the data processing methodology used to create each dataset, including processing, fraud label mapping between the Medicare datasets and the LEIE, and one hot encoding for categorical variables is discussed. The information within each dataset is based on CMS's administrative claims data for Medicare beneficiaries enrolled in the Fee-For-Service program. Note, this data does not take into account any claims submitted through the Medicare Advantage program. Since CMS records all claims information after payments are made, we assume the Medicare data is already cleansed and is correct. Note that NPI is not used in the data mining step, but rather for aggregation and identification. Additionally, for each dataset, we added a year variable which is also used for aggregation and identification.

2.2 Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature

Inappropriate payments by insurance organizations or third party payers occur because of errors, abuse and fraud. The scale of this problem is large enough to make it a priority issue for health systems. Traditional methods of detecting health care fraud and abuse are time-consuming and inefficient. Combining automated methods and statistical knowledge lead to the emergence of a new interdisciplinary branch of science that is named Knowledge Discovery from Databases (KDD). Data mining is a core of the KDD process. Data mining can help third-party payers such as health insurance organizations to extract useful information from thousands of claims and identify a smaller subset of the claims or claimants for further assessment.

We reviewed studies that performed data mining techniques for detecting health care fraud and abuse, using supervised and unsupervised data mining approaches. Most available studies have focused on algorithmic data mining without an emphasis on or application to fraud detection efforts in the context of health service provision or health insurance policy. More studies are needed to connect sound and evidence-based diagnosis and treatment approaches toward fraudulent or abusive behaviors. Ultimately, based on available studies, we recommend seven general steps to data mining of health care claims.

To analyse the results and to make the prediction; we need a significant volume of profile pictures of the active twitter users. Then dividing the twitter profiles of the users into 5 different categories and getting their personality traits using image extraction with big-five validation. We are evaluating the results based on colour, image distribution, image type and diversification, and facial presentation

2.3 A Bug Mining Tool to Identify and Analyse Security Bugs Using Naïve Bayes and TF-IDF

Bug report contains a vital role during software development, However bug reports belongs to different categories such as performance, usability, security etc. This paper focuses on security bug and presents a bug mining system for the identification of security and non-security bugs using the term frequency-inverse document frequency (TF -IDF) weights and Naive Bayes. We performed experiments on bug report repositories of bug tracking systems such as Bugzilla and debugger.

In the proposed approach we apply text mining methodology and TF -IDF on the existing historic bug report database based on the bug's description to predict the nature of the bug and to train a statistical model for manually mislabelled bug reports present in the database. The tool helps in deciding the priorities of the incoming bugs depending on the category of the bugs i.e. whether it is a security bug report or a non-security bug report, using naive Bayes. Our evaluation shows that our tool using TF-IDF is giving better results than the naive Bayes method.

This research paper has two most important endeavours:

• The bottom line of our proposed tool is to retrieve useful information of Bug reports from a BUG TRACKING SYSTEM database using text mining through developing a natural language based on description of the bug reports.

• The comparison of our proposed tool model with the Naive Bayes approach. The results of our tool prove to be fruitful, showing more accuracy than the former approach.

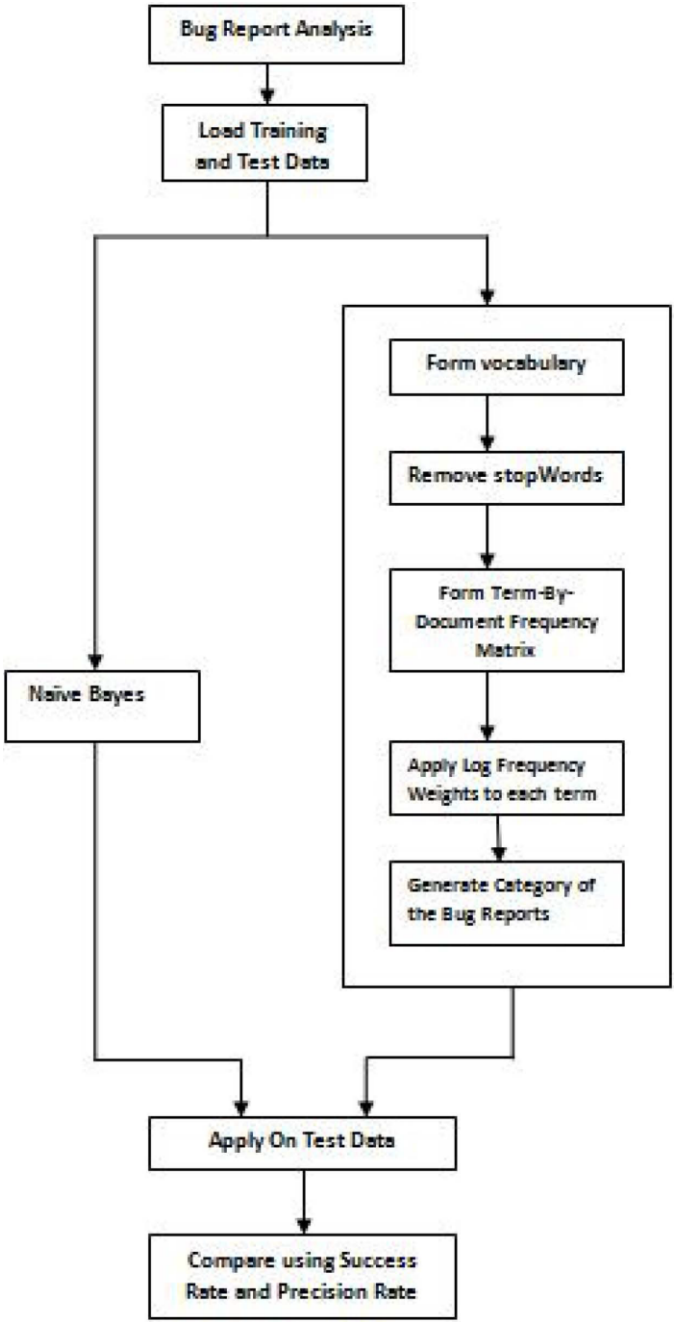

Fig -1: Generalized Algorithm

2.4 Android App Behaviour Classification Using Topic Modelling Technique and Outlier Detection Using App Permission.

Now-a-Days consumption of android apps has become a common phenomenon but user switch from one app to other app is also having high expectancy. There are various causes of Apps' swopping by users. As per research study, one prime reason behind this is that android apps are not providing same functionalities as mentioned in their description on Google Play Store and second crucial reason is that Apps accessing users phone content without taking their permission. The objective of this research work is to classify the apps effectively and identify/detect outlier apps with the help of app behaviour analysis. Outlier apps have been detected to validate whether an android app performs as it claims in its description on Google Play Store as well as other criteria is App accessing user's personal content without user's agreement.

This work has been done in four phases which are as follows- Data extraction phase apps content such as App Title and Description has been crawled and extracted from Google Play Store; Data Pre-processing- this pre-processing phase is required to reduce missing data and high dimension data using filtering and stemming techniques; App classification: formed clusters on the basis of generated feature vector list of various category apps with the help of Topic modelling approaches- probabilistic approach LDA and deterministic approach Non-negative matrix factorization approach NMF; Outlier Detection:- finally for outlier detection used manifest file/ user permission file off apps and mapped its content with App specific features list content to find out outlier Apps.

3. CONCLUSIONS AND FUTURE WORK

In this research we proposed a cost effective fraud detection framework for health care. The detection of frauds in health care is highly challenging task so effective techniques are needed to detect the frauds in this area. Broadly, we classify the fraudulent behaviour into two categories period based claim anomalies and disease based anomalies. The period based claim anomalies are investigated by analysing the statistical decision rules which helps in detecting the outliers and hence frauds and then clustering is performed to simplify the fraud detection process. The disease based anomalies are identified by discovering the association rule mining and identifying the frequent patterns.

The overall objective of research in this area is to get maximum benefit out of medi-claim coverage justifying the investment and reimbursement of claimed services which should be provided. The proposed framework has been evaluated on real world medical data. The performance of the proposed framework is assessed using experimental analysis by involving all the entities like policy providers, policy holders, diseases etc. to get clear of frauds at every level. The results show that our proposed approach is efficient to identify the fraudulent claims from the existing data using data mining techniques.

REFERENCES

[1] Behl, D., Handa, S., & Arora, A. (2014, February). A bug mining tool to identify and analyze security bugs using naïve bayes and tf-idf. In Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on (pp. 294-299). IEEE.

[2] Thornton, Dallas, et al. "Outlier-based health insurance fraud detection for us medicaid data." (2014): 684-694.

[3] 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.

[4] Garg, M., Manga, A., Bhatt, P., & Arora, A. (2016, December). Android app behaviour classification using topic modelling techniques and outlier detection using app permissions. In Parallel, Distributed and Grid Computing (PDGC), 2016 Fourth International Conference on (pp. 500-506). IEEE.

[5] Travaille, Peter, et al. "Electronic fraud detection in the US medicaid healthcare program: lessons learned from other industries." (2011).

[6] 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.