IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

IDENTIFYING HEALTH INSURANCE CLAIM FRAUDS USING MIXTURE OF CLINICAL CONCEPTS

[1]Mrs R. Jhansi rani
[1]Assistant Professor
[1]Department of Computer Applications
[1]Chadalawada Ramanamma Engineering College
(Autonomous),Tirupathi

[2] J. Jaya Prakash
[2] Student
[2] Department of Computer Applications
[2] Chadalawada Ramanamma Engineering College
(Autonomous), Tirupathi

ABSTRACT

Patients depend on health insurance provided by the government systems, private systems, or both to utilize the high-priced healthcare expenses. This dependency on health insurance draws some healthcare service providers to commit insurance frauds. Although the number of such service providers is small, it is reported that the insurance providers lose billions of dollars every year due to frauds. In this article, we formulate the fraud detection problem over a minimal, definitive claim data consisting of medical diagnosis and procedure codes. We present a solution to the fraudulent claim detection problem using a novel representation learning approach, which translates diagnosis and procedure codes into Mixtures of Clinical Codes (MCC). We also investigate extensions of MCC using Long Short Term Memory networks and Robust Principal Component Analysis. Our experimental results demonstrate promising outcomes in identifying fraudulent records.

.

Keywords:- Insurance, Medical diagnostic imaging, Medical services, Encoding, Principal component analysis, Industries, Bayes methods

1.INTRODUCTION

DATA analytics has progressively become crucial to almost any economic development area. Since healthcare is one of the largest financial sectors in the US economy, the massive amount of data, including health records, clinical data, prescriptions, insurance claims, provider information, and patient information “potentially” presents incredible opportunities for data analysts. Health insurance agencies process billions of claims every year and healthcare expenses is over three trillion dollars in the United States .Figure 1 presents a concise flow of a typical healthcare reconciliation process by using different entities involved. First, the service provider’s office ensures that the patient has adequate coverage through his/her insurance plan or other funds before getting any service. Next, the service provider identifies relevant diagnoses based on the initial examinations performed on the patient. The service provider then runs tests on the patient using one or more medical interventions such as further diagnostics and surgical procedures. These diagnoses and procedures are usually tagged with the patient's report along with other information such as personal, demographic, and past/present visit information. At this point, the patient typically pays a copay defined in his/her insurance plan and checks out. Then, the patient's report is sent to a medical coder who abstracts the information and creates a "superbill" containing all information about the provider, Given the economic volume of the healthcare industry, it is natural to observe fraudulent and fabricated claims submitted to insurance companies. The National Health Care Anti- Fraud Association (NHCAA) defines healthcare fraud as "An intentional deception or misrepresentation made by a person, or an entity, with the knowledge that the deception could result in some unauthorized benefit to him or some other entities" Those fabricated claims bear a very high cost, albeit they constitute a small fraction. According to NHCAA

the fraud related financial loss is in the orders of tens of billions of dollars in the United States. Although there are strict policies regarding fraud and abuse control in healthcare industries, studies show that a very small portion of the losses are recovered annually.

Most typical fraudulent activities committed by dishonest providers in the healthcare domain include the following.

_ Making false diagnoses to justify procedures that are not medically necessary.

_ Billing for high priced procedures or services instead of the actual procedures, also called "upcoding".

_ Fabricating claims for unperformed procedures.

_ Performing medically unnecessary procedures to claim insurance payments.

_ Billing for each step of a procedure as if it is a separate procedure, also called "unbundling".

_ Misrepresenting non-covered treatments as medically necessary to receive insurance payments, especially for cosmetic procedures.

It is not feasible or practical to apply only domain knowledge to solve all or a subset of the issues listed above. Automated data analytics can be employed to detect fraudulent claims at an early stage and immensely help domain experts to manage the fraudulent activities much better.

In this paper, we focus on the problem of healthcare fraud detection from health insurance providers' viewpoint. We answer the question of how to classify a procedure as legitimate or fraudulent from a claim when we only have limited data available, i.e. diagnosis and procedure codes. The problem of fraud detection in medical domain has been identified using different approaches such as data mining classification methods Bayesian analysis statistical surveys non-parametric approaches, and expert analysis. Existing methods use physicians profile, background history, claim amount, service quality, services performed per provider, and related metrics from a claim database to create models for claim status prediction. Although these methods are successful, they often employ datasets that are not publicly available. Furthermore, the variables featured in those datasets are diverse and generally incompatible, which makes the solutions very difficult to transfer. In this study we limit our available data to diagnosis and procedure codes, because obtaining third-party access to richer datasets is often prohibited by Health Insurance Portability and Accountability Act (HIPAA) in the US, General Data Protection Regulation (GDPR) in Europe or similar law in other regions. Besides, the healthcare industry is more apprehensive to share data compared to other sectors. Moreover, different software systems report different patient variables,which prohibits transferring solutions from one system to another. As a result, we confine our problem formulation to diagnosis and procedure codes which can always be handled in the same way whether they are country-specific or international. Our solution approach assumes the claim data as a mixture of medical concepts with respect to clinical codes of diagnoses and procedures in International Classification of Diseases (ICD) coding format. Moreover, the proposed approach works on other coding formats, e.g., Current Procedural Terminology (CPT) and Healthcare Common Procedure Coding System (HCPCS), or their combinations without any modification.

We represent an insurance claim as a Mixture of latent Clinical Concepts (MCC) using probabilistic topic modeling. To the best of our knowledge this is the first work representing insurance claims as mixtures of clinical concepts in a latent space. We assume that every claim is a representation of latent or obvious mixtures of clinical concepts such as pain, mental or infectious diseases. Moreover, each clinical concept is a mixture of clinical codes, i.e., diagnosis and procedure codes. The intuition behind our model comes from the services provided by doctor's offices, clinics, and hospitals. In general, a patient gets services based on specific issues consisting of one or more diagnoses. Next, the service provider performs necessary procedures to treat the patient. Therefore, the diagnoses and procedures in a claim can be represented as a mixture of clinical concepts such as pain, mental, infectious diseases and/or their treatments. Note that, we do not explicitly label or interpret these concepts, as they are often not obvious, complex or require domain knowledge.

We extend the MCC model using Long-Short Term Memory networks and Robust Principal Component Analysis. Our goal in extending MCC is to filter the significant concepts from claims and classify them as fraudulent or non-fraudulent. We extend MCC by using the concept weights of a claim as a sequence representation within a Long-Short Term Memory (LSTM) network. This network allows us to represent the claims as sequences of dependent concepts to be classified by the LSTM. Similarly, we apply Robust Principal Component Analysis (RPCA) to filter significant concept weights by decomposing claims into a low-rank and sparse vector representations. The low-rank matrix ideally captures the noise-free weights.

Our unique contributions in this study can be summarized as follows.

_ We formulate the fraudulent claim detection problem over a minimal, definitive claim data consisting of procedure and diagnosis codes.

_ We introduce clinical concepts over procedure and diagnosis codes as a new representation learning approach.

_ We extend the mixtures of clinical concepts using LSTM and RPCA for classification.

We compare our approaches to the Multivariate Outlier Detection (MOD) [11] and a baseline method and report improved performance. Multivariate Outlier Detection method consists of two steps which are used to detect anomalous provider payments within Medicare claims data. In the first step, a multivariate regression model is built on 13 hand picked features to generate corresponding residuals. Next, the residuals are used as inputs to a generalized univariate probability model. Specifically, they used probabilistic programming methods in Stan to identify possible outliers in the claim data. The authors use the same CMS (Centers for Medicare and Medicaid Services) dataset that we use in our experiments with a different problem formulation. Their study incorporates providers and beneficiary data that was related to Medicare beneficiaries within the state of Florida, while we employ MOD on MCC features. On the other hand, the baseline classifier assigns a test claim as the majority label present in the training claim data.

Our experimental results show that MCC + LSTM reaches an accuracy, precision, and recall scores of 59%, 61%, and 50%, respectively on the inpatient dataset obtained from CMS. In addition, it demonstrates 78%, 83%, and 72% accuracy, precision, and recall scores, respectively on the outpatient dataset We believe that the proposed problem formulation, representation learning and solution will initiate new research on fraudulent claim detection using minimal, but definitive data. The rest of the paper is organized as follows. Section II presents the related work. We formally introduce the problem and present our solution in Section III. Section IV demonstrates the empirical evaluations. Finally, we conclude the paper in Section V.

2. LITERATURE SURVEY

2.1 DIFFERENT AUTHORS DISCUSSION:

It is not feasible or practical to apply only domain knowledge to solve all or a subset of the issues listed above. Automated data analytics can be employed to detect fraudulent claims at an early stage and immensely help domain experts to manage the fraudulent activities much better.

2.2 DOMAIN DESCRIPTION:

The fraud related financial loss is in the orders of tens of billions of dollars in the United States. Although there are strict policies regarding fraud and abuse control in healthcare industries, studies show that a very small portion of the losses are recovered annually.

3. PROBLEM STATEMENT

3.1 EXISTING SYSTEM

Yang and Hwang developed a fraud detection model using the clinical pathways concept and process-mining framework that can detect frauds in the healthcare domain [13]. The method uses a module that works by discovering structural patterns from input positive and negative clinical instances. The most frequent patterns are extracted from every clinical instance using the module. Next, a feature-selection module is used to create a filtered dataset with labeled features. Finally, an inductive model is built on the feature set for evaluating new claims. Their method uses clustering, association analysis, and principal component analysis. The technique was applied on a real-world data set collected from National Health Insurance (NHI) program in Taiwan. Although the authors constructed different features to generate patterns for both normal and abusive claims, the significance of those features is not discussed.

Bayerstadler presented a predictive model to detect fraud and abuse using manually labeled claims as training data. The method is designed to predict the fraud and abuse score using a probability distribution for new claim invoices. Specifically, the authors proposed a Bayesian network to summarize medical claims' representation patterns using latent variables. In the prediction step, a multinomial variable modeling predicts the probability scores for various fraud events. Additionally, they estimated the model parameters using Markov Chain Monte Carlo (MCMC) .

Zhang proposed a Medicare fraud detection framework using the concept of anomaly detection .First part of the proposed method consists of a spatial density based algorithm which is claimed to be more suitable compared to local outlier factors in medical insurance data. The second part of the method uses regression analysis to identify the linear dependencies among different variables. Additionally, the authors mentioned that the method has limited application on new incoming data

Kose used interactive unsupervised machine learning where expert knowledge is used as an input to the system to identify fraud and abuse related legal cases in healthcare. The authors used a pairwise comparison method of analytic hierarchical process (AHP) to incorporate weights between actors (patients) and attributes. Expectation maximization (EM) is used to cluster similar actors. They had

domain experts involved at different levels of the study and produced storyboard based abnormal behavior traits. The proposed framework is evaluated based on the behavior traits found using the storyboard and later used for prescriptions by including all related persons and commodities such as drugs.

Bauder and Khoshgoftaar proposed a general outlier detection model using Bayesian inference to screen healthcare claims. They used Stan model which is similar to in their experiments. Note that, they consider only provider level-fraud detection without considering clinical code based relations. Many of those methods use private datasets or different datasets with incompatible feature lists. Therefore, it is very difficult to directly compare these studies. In addition, HIPAA, GDPR and similar law enforce serious penalties for violations of the privacy and security of healthcare information, which make healthcare providers and insurance companies very reluctant to share rich datasets if not at all. For these reasons, we formulate the problem over a minimal, definitive claim data consisting of diagnosis and procedure codes. Under this setting we tackle the problem of flagging a procedure as legitimate or fraudulent using mixtures of clinical codes along with RNN and RPCA based encodings.

3.2 DISADVANTAGE OF EXISTINTG SYSTEM:

Making false diagnoses to justify procedures that are not medically necessary. Fabricating claims for unperformed procedures. Performing medically unnecessary procedures to claiminsurance payments. Billing for each step of a procedure as if it is a separateprocedure, also called “unbundling”. Misrepresenting non-covered treatments as medicallynecessary to receive insurance payments, especially forcosmetic procedures.

4.PROPOSED SYSTEM

4.1 PROPOSED SYSTEM:

We extend the MCC model using Long-Short Term Memory networks and Robust Principal Component Analysis. Our goal in extending MCC is to filter the significant concepts from claims and classify them as fraudulent or non-fraudulent. We extend MCC by using the concept weights of a claim as a sequence representation within a Long-Short Term Memory (LSTM) network. This network allows us to represent the claims as sequences of dependent concepts to be classified by the LSTM. Similarly, we apply Robust Principal Component Analysis (RPCA) to filter significant concept weights by decomposing claims into a low-rank and sparse vector representations. The low-rank matrix ideally captures the noise-free weights.

Our unique contributions in this study can be summarized as follows.

The system formulates the fraudulent claim detection problem over a minimal, definitive claim data consisting of procedure and diagnosis codes.

The system introduces clinical concepts over procedure and diagnosis codes as a new representation learning approach.

The system extends the mixtures of clinical concepts using LSTM and RPCA for classification.

4.2 ADVANTAGE OF PROPOSED SYSTEM:

The proposed system uses Support Vector Machine (SVM) for classification with MCC. Multivariate Outlier Detection method is an effective method which is used to detect anomalous provider payments within Medicare claims data.

5.IMPLEMENTATION

5.1 Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Login, Browse and Train & Test Health Insurance Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Prediction Of Health Insurance Fraud Type, View Health Insurance Fraud Type Ratio, Download Predicted Data Sets, View Health Insurance Fraud Type Ratio Results, View All Remote Users

5.2 View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

5.3 Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT HEALTH INSURANCE CLAIM FRAUD TYPE, VIEW YOUR PROFILE.

6.CONCLUSION

In this paper, we pose the problem of fraudulent insurance claim identification as a feature generation and classification process. We formulate the problem over a minimal, definitive claim data consisting of procedure and diagnosis codes, because accessing richer datasets are often prohibited by law and present inconsistencies among different software systems. We introduce clinical concepts over procedure and diagnosis codes as a new representation learning approach. We assume that every claim is a representation of latent or obvious Mixtures of Clinical Concepts which in turn are mixtures of diagnosis and procedure codes. We extend the MCC model using Long-Short Term Memory network (MCC + LSTM) and Robust Principal Component Analysis (MCC + RPCA) to filter the significant

concepts from claims and classify them as fraudulent or non fraudulent. Our results demonstrate an improvement scope to find fraudulent healthcare claims with minimal information. Both MCC and MCC + RPCA exhibit consistent behavior for varying concept sizes and replacement probabilities in the negative claim generation process. MCC + LSTM reaches an accuracy, precision, and recall scores of 59%, 61%, and 50%, respectively on the inpatient dataset. Besides, it presents 78%, 83%, and 72% accuracy, precision, and recall scores, respectively on the outpatient dataset. We notice similarity between the results of MCC and

MCC + RPCA, as both use an SVM classifier. We believe that the proposed problem formulation, representation learning and solution will initiate new research on fraudulent insurance claim detection using minimal, but definitive data.

7. FUTURE ENHANCEMENT

The system formulates the fraudulent claim detection problem over a minimal, definitive claim data consisting of procedure and diagnosis codes.

8.REFERENCES

National Health Care Anti-Fraud Association, "The challenge of health care fraud," https://www.nhcaa.org/resources/health-care-antifraud- resources/the-challenge-of-health-care-fraud.aspx, 2020, accessed January, 2020.

Font Awesome, "Image generated by free icons," https://fontawesome.com/license/free, 2020, online.

National Health Care Anti-Fraud Association, "Consumer info and action," https://www.nhcaa.org/resources/health-care-anti-fraudresources/ consumer-info-action.aspx, 2020, accessed January, 2020.

W. J. Rudman, J. S. Eberhardt, W. Pierce, and S. Hart-Hester, "Healthcarefraud and abuse," Perspectives in Health Information Management/ AHIMA, American Health Information Management Association, vol. 6, no. Fall, 2009.

M. Kirlidog and C. Asuk, "A fraud detection approach with data mining in health insurance," Procedia-Social and Behavioral Sciences, vol. 62, pp. 989–994, 2012.

V. Rawte and G. Anuradha, "Fraud detection in health insurance using data mining techniques," in 2015 International Conference on Communication, Information & Computing Technology (ICCICT)., 2015, pp. 1–5.

C. Phua, D. Alahakoon, and V. Lee, "Minority report in fraud detection: classification of skewed data," Acm sigkdd explorations newsletter, vol. 6, no. 1, pp. 50–59, 2004.

T. Ekina, F. Leva, F. Ruggeri, and R. Soyer, "Application of Bayesian methods in detection of healthcare fraud," chemical engineering Transaction, vol. 33, 2013.

J. Li, K.-Y. Huang, J. Jin, and J. Shi, "A survey on statistical methods for health care fraud detection," Health care management science, vol. 11, no. 3, pp. 275–287, 2008.

R. J. Freese, A. P. Jost, B. K. Schulte, W. A. Klindworth, and S. T.Parente, "Healthcare claims fraud, waste and abuse detection system using non-parametric statistics and probability based scores," Jan. 19 2017, uS Patent App. 15/216,133.

R. A. Bauder and T. M. Khoshgoftaar, "Multivariate anomaly detection in medicare using model residuals and probabilistic programming," in The Thirtieth International Flairs Conference, 2017.

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," Journal of statistical software, vol. 76, no. 1, 2017.

W.-S. Yang and S.-Y. Hwang, "A process-mining framework for the detection of healthcare fraud and abuse," Expert Systems with Applications, vol. 31, no. 1, pp. 56–68, 2006.

A. Bayerstadler, L. van Dijk, and F.Winter, "Bayesian multinomial latent variable modeling for fraud and abuse detection in health insurance," Insurance: Mathematics and Economics, vol. 71, pp. 244–252, 2016.

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, "Introducing markov chain monte carlo," Markov chain Monte Carlo in practice, vol. 1, p. 19, 1996.

W. Zhang and X. He, "An anomaly detection method for medicarefraud detection," in Big Knowledge (ICBK), 2017 International Conference on., 2017, pp. 309–314.

L. Zhang, J. Lin, and R. Karim, "Adaptive kernel density-based anomaly detection for nonlinear systems," Knowledge-Based Systems, vol. 139, pp. 50–63, 2018.

I. Kose, M. Gokturk, and K. Kilic, "An interactive machine-learningbased electronic fraud and abuse detection system in healthcare insurance," Applied Soft Computing, vol. 36, pp. 283–299, 2015.

R. A. Bauder and T. M. Khoshgoftaar, "A probabilistic programming approach
for outlier detection in healthcare claims," in Machine Learning and Applications (ICMLA), 2016 15th International Conference on., 2016, pp. 347–354.

J. Wang and S. Luo, "Augmented beta rectangular regression models: A bayesian perspective," Biometrical Journal, vol. 58, no. 1, pp. 206–221, 2016.

Centers for Medicare and Medicaid Services, "ICD-10," https://www.cms.gov/Medicare/Coding/ICD10/, 2020, accessed January, 2020.

Medical Billing and Coding, "HCPCS codes," https://www.medicalbillingandcoding.org/hcpcs-codes/, 2020, accessed January, 2020.

American Academy of Professional Coders, "CPT codes," https://coder.aapc.com/cpt-codes, 2020, accessed January, 2020.

S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, vol. 3, no. Jan, pp. 993–1022, 2003.