Predicting Healthcare Insurance Fraud Using Markov Observation Models

Abhimanyu Nag

February 2024

1 Introduction

In 2011, A federal jury in Baltimore convicted cardiologist John R. McLean, of Salisbury, Maryland, on six health care fraud offences in connection with a scheme in which Dr. McLean submitted insurance claims for inserting unnecessary cardiac stents, ordered unnecessary tests, and made false entries in patient medical records, in order to defraud Medicare, Medicaid and private insurers, raking in a whopping $711,583 [1]. Health care fraud cases like these have been on the rise for the past few years and the costs associated with fraud are passed on to the population in the form of increased premiums or serious harm to beneficiaries. There is an intense need for digital healthcare fraud detection systems to evolve in combating this societal threat. In this spirit, we introduce a novel statistical approach, developed by Professor Kouritzin[3] at the University of Alberta to model the simulated health insurance claims in a sequential set up using counting to capture fraudulent claims whenever they appear. It is a Hidden Markov Model which is expanded to allow for Markov chain observations called Markov Observation Model (MOM) that would estimate the most likely sequence of hidden states given the sequence of observations with high accuracy. The MOM has shown tremendous results when it comes to predicting outcomes and patterns, with an impressive 93% accuracy in predicting coin flips, as compared to other methods which have had an accuracy between 60% to 85%.

2 Overview of Concepts

2.1 Hidden Markov Models

A Hidden Markov Model (HMM) is a statistical model used to describe a system that evolves over time and produces a sequence of observable outputs. It is a probabilistic model comprising two interrelated stochastic processes: a hidden process (denoted as X) and an observable process (denoted as Y). The observations Y are dependent on the hidden states X in a known manner, but

the states of X are not directly observable. The main objective of an HMM is to infer or estimate the sequence of hidden states X based on the observed sequence Y. The model satisfies the Markov property, where the current state of Y depends only on the current state of X, and the future state of X depends only on the current state of X. HMMs find applications in various fields such as computational finance, biological sequence analysis, speech recognition, and natural language processing. More details found in the original papers by Baum and collaborators [2]

2.2 Markov Observation Models

Markov Observation Models (MOM) are an extension of traditional Hidden Markov Models (HMMs) that allow for observations to be a Markov chain, addressing the limitation of HMMs. In MOM, the hidden state is a homogeneous Markov chain (i.e constant transition probability) , and the observations are also assumed to be a Markov chain whose one-step transition probabilities depend on the hidden Markov chain. This model allows us to estimate transition probabilities for both the hidden state and the observations, as well as the initial joint hidden-state-observation distribution. MOM aims to narrow the gap between the limited HMM and more complex models while still allowing an extension of the Baum Welch algorithm and a modified Viterbi Algorithm.

3 Research Problem

3.1 Research Goal

Our research question is this : Can we use the Markov Observation Model (MOM) on a large scale healthcare insurance dataset to quantify and predict fraudulent insurance claims and even model mindsets of the exact perpetrators of the crime? A variety of literature exist on this subject (see [4, 5]) which employ techniques such as outlier prediction and big data analytics. We aim for our implementation to be the first motivation of the use of Hidden Markov Model approach to predict healthcare fraud and other problems of the same flavour.

3.2 Proposed Methodology

We employ the use of a canned dataset using a mixture of real patient and simulated data (in order to maintain privacy of patients). The dataset has 9 columns and 636000 data-points. Each column represents a different metric to describe the patient, the medical service provider and the claim amounts submitted by the service providers. The methodology of research involves the following :

1. We define a set U of marks representing possible transactions, such as

service types, prescriptions, and dollar amounts. We partition all transactions into bins, ensuring that each transaction fits into exactly one bin.

2. We define event types E_i for each claim, and create observations Y_t representing the occurrences of each mark type up to time t. This is done by counting the events with mark type u over time.

3. We construct a canonical model $q_{y \to y'}$ by considering all providers together. We also assume a Bernoulli Counting Process for probability increments $q_{y \to y'}$.

4. We utilize the EM algorithm to model each provider's behavior, determining probability parameters with a large number of hidden states. Compare these provider-specific models to the canonical model using Bayesian methods, focusing on periods with the highest Bayes' factor increase, termed deviant states.

5. At last, we employ the MOM prediction, starting from each deviant state, to predict fraud based on price changes associated with the marks.

4 Expected Results and Future Work

We expect the Markov Observation Model to tell us which service providers have been defrauding insurance providers through modelling the mindsets. This would imply a great feat in the usefulness of the MOM both theoretically and practically. This only opens a box of questions about the effectiveness of the MOM in other related problems. We aim to motivate and popularize the use of this algorithm in more areas such as network security and economic modelling which can be undertaken as part of graduate study in the future.

References

[1] https://archives.fbi.gov/archives/baltimore/press-releases/2011/salisbury-cardiologist-convicted-of-implanting-unnecessary-cardiac-stents

[2] Baum, L. E. and Petrie, T. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. The Annals of Mathematical Statistics. 37 (6): 1554-1563. doi:10.1214/aoms/1177699147.

[3] Kouritzin, Michael A. "Markov Observation Models." *arXiv preprint arXiv:2208.06368* (2022).

[4] Dora, Prajna and Dr. G. Hari Sekharan. "Healthcare Insurance Fraud Detection Leveraging Big Data Analytics." (2015).

[5] Anbarasi, M. and S. Dhivya. "Fraud detection using outlier predictor in health insurance data." *2017 International Conference on Information Communication and Embedded Systems (ICICES)* (2017): 1-6.