Fraud Detection in Health Insurance using Data Mining Techniques

Vipula Rawte
Student, M.E. (Computer Engineering)
St.Francis Institute of Technology
Mumbai-400103, India.
rawtevipula25@gmail.com

G Anuradha
Associate Professor (Computer Engineering)
St.Francis Institute of Technology
Mumbai-400103, India.
ganusrinu4@yahoo.co.in

***Abstract*—Fraud is widespread and very costly to the health-care insurance system. Fraud involves intentional deception or misrepresentation intended to result in an unauthorized benefit. It is shocking because the incidence of health insurance fraud keeps increasing every year. In order to detect and avoid the fraud, data mining techniques are applied. This includes some preliminary knowledge of health care system and its fraudulent behaviors, analysis of the characteristics of health care insurance data. Data mining which is divided into two learning techniques viz., supervised and unsupervised is employed to detect fraudulent claims. But, since each of the above techniques has its own set of advantages and disadvantages, by combining the advantages of both the techniques, a novel hybrid approach for detecting fraudulent claims in health insurance industry is proposed.**

Keywords—data mining; health insurance fraud; supervised; unsupervised

I. INTRODUCTION

Deliberately deceiving the health insurance company that results in healthcare benefits being paid illegitimately to an individual or group is known as health insurance fraud. The main purpose of fraud is financial benefit. According to a recent survey, it is estimated that the number of false claims in the industry is approximately 15 per cent of total claims. Insurance companies in USA incur losses over 30 billion USD annually to healthcare insurance frauds. The statistics is appalling in developing country like India as well. The report suggests that the healthcare industry in India is losing approximately Rs 600-Rs 800 crores incurred on fraudulent claims annually [1]. Frauds blow a hole in the insurance industry. Health insurance is a bleeding sector with very high claims ratio. So, to make health insurance industry free from fraud, it is necessary to focus on elimination or minimization of fake claims arriving through health insurance.

The health insurance fraud claims are broadly classified under the following headings:

- Billing for services not rendered: Billing insurance company for things that never happened. Example: Forging the signature of those involved in giving bills.
- Upcoding of services: Billing insurance company for services that are costlier than the actual procedure that was done. Example: 45-minute session being billed as 60-minute session
- Upcoding of items: Billing insurance company for medical equipment that is costlier than the actual equipment. Example: Billing for power assisted wheelchair while giving the patient only the manual wheelchair.
- Duplicate claims: Not submitting exactly the same bill, but changing some small portion like the date in order to charge insurance company twice for the same service rendered. Example: An exact copy of the original claim is not filed for the second time, but rather some portion like date is changed to get the benefit twice the original.
- Unnecessary services: Filing claims which in no way apply to the condition of a patient. Example: Patient with no symptoms of diabetes filing claim for daily usage of insulin injections.

II. DATA MINING

Nowadays there is huge amount of data stored in real-world databases and this amount continues to grow fast. So, there is a need for semi-automatic methods that discover the hidden knowledge in such database. Data mining automatically filtering through immense amounts of data to find known/unknown patterns, bring out valuable new perceptions and make predictions.

Data mining techniques tend to learn models from data. There are two approaches on learning the data mining models. Those are supervised learning, unsupervised learning; and they are described below:

A. Supervised Learning:

This is the most usual learning technique wherein the model is trained using pre-defined class labels. In the context of health insurance fraud detection the class labels may be the "legitimate" and "fraudulent" claims. The training dataset can

978-1-4799-5522-0/15/$31.00 ©2015 IEEE