IJCSN International Journal of Computer Science and Network, Volume 5, Issue 2, April 2016
ISSN (Online) : 2277-5420 www.IJCSN.org
Impact Factor: 1.02

Fraud Claim Detector for Health Insurance

[1] **Aiswarya C Pradeep,** [2] **Aswathy P.,** [3] **Hiba Sakkir,** [4] **Magna A,** [5] **Naseeha T K**

[1,2,3,4,5] Student, Dept. of Computer Science
MESCE, Kuttippuram.

Abstract - A health insurance system is an organization that provide health care services to meet the health needs of target population. Social health insurance is one of the primary methods of funding health insurance systems. Fraud is very rampant in today's society and results in huge loss for the health insurance system. Intentional deception and misrepresentation are involved in frauds, so it will result in unauthorized benefits. Data mining is the technique applied to detect the fraudulent claims in the health insurance system. Data mining is basically a filtering through a large amount of data to get useful information. Elimination of fake claims is necessary to make health insurance industry free fraud.

Keywords - *Data Mining, SVM, Unsupervised, Supervised.*

1. Introduction

Filling of dishonest health care claims in order to get a profit is known as health care fraud. Money benefit is the main agenda of fraud. According to a recent study, 15 percentage of total claims are fraud in the industry. Due to the spread of fraud, every dollar spent on health care goes towards paying for fraudulent health care claims.The health care industry in India is losing nearly Rs. 600-Rs 800 crores due to fraudulent claims. A health insurance system should pay a legitimate claim within 30 days. However because of the 30-day rule, agencies dont have enough time to perform an adequate investigation before an insurance has to pay. Fraud claims are of different types:

- Altering small portion of the bill for charging insurance company twice for same purpose
- Charging insurance company for services not happened.
- Loaning another insurance card.

Health insurance fraud differs from a lot of other insurance fraud, in that it is not normally perpetrated by the health insurance policy holder him or herself. Normally, health insurance fraud will be perpetrated by the health care provider, instead, whether through fraudulent billing practices or unnecessary prescriptions and referrals. Health insurance fraud is also problematic in consideration, if only because many doctors often perpetrate health insurance fraud out of a desire to help their patients. They will bill the insurance company for a different operation from the one they did, so that the operation billed will be covered under the patient's health insurance policy, for instance. Such actions are still considered health insurance fraud, even though they are performed with the best of intentions. But standard personal gain oriented fraud exists in the domain of health insurance, as well, with many versions of fraud that will result in the fraudster physician being able to charge more, by "upgrading" his charges, or by performing unnecessary services. To find out more about health insurance fraud and the ways in which doctors may perpetrate it, click the link.

Health insurance is a form of insurance that pays for medical expenses. If you are covered under health insurance, you pay some amount of premium every year to an insurance company and if you have an accident or if you have to undergo an operation or a surgery, the insurance company will pay for the medical expenses. With health insurance providing a world of benefits to people, fraudulent claims are on the rise.Frauds can be committed by anybody. It can be committed by a policyholder, a health insurance company or even its employees.Frauds committed by a policyholder could consist of members that are not eligible, concealment of age, concealment of pre-existing diseases, failure to report any vital information, providing false information regarding self or any other family member, failure in disclosing previously settled or rejected claims, frauds in physicians prescriptions, false documents, false bills, exaggerated claims, etc.

2. Data Mining

The data in real-world database continues to grow fast and hence we require to handle a huge amount of data. So data

mining techniques are used to discover hidden useful knowledge in such a database. Data mining basically filters through huge amount of data to make required perceptions and predictions. Data mining is an integral part of knowledge discoverydatabases, which is the overall process of converting raw data into useful information.

The two learning approaches in data mining are:

- Supervised learning

In this learning technique, model is trained by a pre-defined class labels. Here the class labels are legitimate and fraudulent claims. When a new claim is arrive, it is compared with this pre-defined class. If it is similar to the fraudulent class then it is classified as illegitimate. It can be used for pattern classification easily. But its main disadvantage is it cannot detect new type frauds.

In order to solve a given problem of supervised learning, one has to perform the following steps:

- Determine the type of training examples. Before doing anything else, the user should decide what kind of data is to be used as a training set. In the case of handwriting analysis, for example, this might be a single handwritten character, an entire handwritten word, or an entire line of handwriting.
- Gather a training set. The training set needs to be representative of the real-world use of the function. Thus, a set of input objects is gathered and corresponding outputs are also gathered, either from human experts or from measurements.
- Determine the input feature representation of the learned function. The accuracy of the learned function depends strongly on how the input object is represented. Typically, the input object is transformed into a feature vector, which contains a number of features that are descriptive of the object. The number of features should not be too large, because of the curse of dimensionality; but should contain enough information to accurately predict the output.
- Determine the structure of the learned function and corresponding learning algorithm. For example, the engineer may choose to use support vector machines or decision trees.
- Complete the design. Run the learning algorithm on the gathered training set. Some supervised learning algorithms require the user to determine certain control parameters. These parameters may be adjusted by optimizing performance on a subset (called a validation set) of the training set, or via cross-validation.
- Evaluate the accuracy of the learned function. After parameter adjustment and learning, the performance of the resulting function should be measured on a test set that is separate from the training set.
- Unsupervised Learning

There is no pre-defined class labels, so that it is not restricted to a particular pattern. It can detect both new and old fraud claims.

- Approaches for learning latent variable models such as
- Expectation-maximization algorithm (EM)
- Method of moments
- Blind signal separation techniques

Consider a machine (or living organism) which receives some sequence of inputs x1, x2, x3. . . where xt is the sensory input at time t. This input, which we will often call the data, could correspond to an image on the retina, the pixels in a camera, or a sound waveform. It could also correspond to less obviously sensory data. In unsupervised learning the machine simply receives inputs x1, x2. . . but obtains neither supervised target outputs, nor rewards from its environment. It may seem somewhat mysterious to imagine what the machine could possibly learn given that it doesn't get any feedback from its environment. However, it is possible to develop of formal framework for unsupervised learning based on the notion that the machine's goal is to build representations of the input that can be used for decision making, predicting future inputs, efficiently communicating the inputs to another machine, etc. In a sense, unsupervised learning can be thought of as finding patterns in the data above and beyond what would be considered pure unstructured noise.

3. Literature Survey

There are many data mining techniques out of which the following are chosen.

3.1 Anomaly Detection

In this method, probability of a claim to be fraudulent is calculated by reviewing the previous claims.In data mining, anomaly detection (or outlier detection) is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset. Typically the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or errors in a text. Anomalies are

also referred to as outliers, novelties, noise, deviations and exceptions.

In particular in the context of abuse and network intrusion detection, the interesting objects are often not rare objects, but unexpected bursts in activity. This pattern does not adhere to the common statistical definition of an outlier as a rare object, and many outlier detection methods (in particular unsupervised methods) will fail on such data, unless it has been aggregated appropriately. Instead, a cluster analysis algorithm may be able to detect the micro clusters formed by these patterns.

Three broad categories of anomaly detection techniques exist. Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal by looking for instances that seem to fit least to the remainder of the data set. Supervised anomaly detection techniques require a data set that has been labeled as ”normal” and ”abnormal” and involves training a classifier (the key difference to many other statistical classification problems is the inherent unbalanced nature of outlier detection). Semi-supervised anomaly detection techniques construct a model representing normal behavior from a given normal training data set, and then testing the likelihood of a test instance to be generated by the learnt model.

3.2 Support Vector Machines

A decision boundary is determined between classes of legitimate and fraudulent claims for the purpose of classification. Each insurance claim is then placed into either legitimate or fraudulent class.In machine learning, support vector machines (SVMs, also support vector networks[1]) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

When data are not labeled, a supervised learning is not possible, and an unsupervised learning is required, that would find natural clustering of the data to groups, and map new data to these formed groups. The clustering algorithm which provides an improvement to the support vector machines is called support vector clustering[2] and is often used in industrial applications either when data is not labeled or when only some data is labeled as a preprocessing for a classification pass.

4. Proposed System

In the proposed system, claims are clustered using ECM method and then the claims are classified using the SVM method. In the present existing system, there is no mechanism to detect fake claims and this will result in the huge wastage of insurance money.

4.1. Evolving Clustering Method

Evolving clustering method (ECM) is used to cluster data. Clustering means grouping data of similar properties. Each cluster have a radius that decide the cluster boundary. Initially radius is set to zero and it will increase by adding data. ECM comes under unsupervised learning. The ECM is a fast, onepass algorithm for dynamic clustering of an input stream of data. It is a distance based clustering method where the cluster centers are represented by evolved nodes in an on-line mode. This is based on the concept ofmadding and modifying the clusters as new data is presented, where the modification to the clusters affects both the position of the clusters and the size of the cluster, in terms of a radius parameter associated with each cluster that determines the boundaries of that cluster. ECM has only one parameter, which drives the addition of clusters, known as the distance threshold. When new clusters are added, their centers are set to equal the example that triggered their creation, and the radius R of a new cluster is initially set to zero. R grows as more vectors are allocated to the cluster.

4.2. Support Vector Machines

This technique is a supervised learning techniques used in classification. In the initial phase data already classified is fed to the algorithm. After this, SVM predicts the class of the incoming data.

Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships.

4.3. Block Diagram

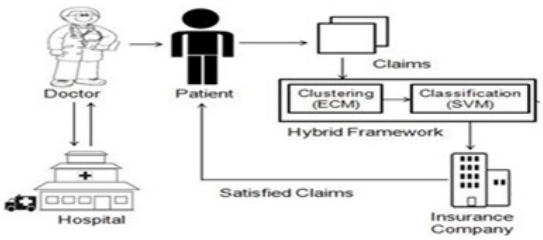

Fig. 1. Block diagram

Steps involved :

- Hospitals are registered publically and insurance company will approve or reject
- Doctors are registered by hospital. Registered doctors bill patients for services rendered.
- Claims are filed to the insurance company.
- Insurance company pays legitimate claims to patients.

4.4. Pseudo Code for Hybrid Model

- Clustering according to disease type is done by applying ECM for the incoming health insurance claim.
- Classification of clusters into fraud and legitimate classes are done by applying SVM.

Go to clustering step to cluster new claims and repeat.

5. Conclusion

Fraud becomes more diverse, as the amount of data grows. Reduction of fraud can result by elimination of fake claims. Data mining finds useful hidden patterns to present the required knowledge. After considering the advantages and disadvantages of most of the classification and clustering techniques. ECM is considered as the clustering technique and SVM as the classification technique. A health insurance claim prepared with the intention to deceive, conceal or distort relevant information that eventually accounts for health care benefits for an individual or a particular group is defined as fraudulent health insurance claim.Frauds by health insurance companies or its employees include preparation of bogus claims by fake physicians, billing for products or services not rendered, exaggerated claims submission, billing prepared for higher level of services, modifications or alterations made in submission of health insurance claims, change in diagnosis of the patient, fake documentation, and fraud committed by the employees of a hospital or any other healthcare product or service provider in order to make a quick buck.Fraudulent and dishonest health insurance claims are a major morale and moral hazard not only for the health insurance industry but even for the entire nation's economy. Concrete proof as evidence including documentation, statements made by the policyholder and his family members and even neighbors are taken into consideration. The essential components of fraud include intention to deceive, derive benefits from the health insurance industry, preparation of exaggerated or inflated claims or medical bills and an intention to induce the firm to pay more than it otherwise would. Devising innovative methods and tactics including pressure tactics, favoritism and nepotism form a part of fraud which is a hazard growing by leaps and bounds since the last decade.To establish that a fraud has been committed requires furnishing of relevant proof. An indepth analysis of the health insurance policyholders intention may also be taken into consideration. It is estimated that the number of false health insurance claims in the industry is approximately 15 per cent of total claims. The report suggests that the healthcare industry in India is losing approximately Rs. 600 to Rs. 800 crores incurred on fraudulent claims annually. Health insurance is a bleeding sector with very high claims ratio. Hence, in order to make health insurance a viable sector, it is essential to concentrate on elimination or minimization of fake claims.

References

[1] Dr. BiswenduBardhan. Fraud in Health Insurance?, http://healthcare.financialexpress.com/200711/market13.shtml.

[2] MelihKirlidoga, CuneytAsuk(2012) A fraud detection approach with data mining in health insurance. Procedia - Social and Behavioral Sciences 62 (2012) 989 ? 994.

[3] Dan Ventura. Class Lecture, Topic: "SVM Example". BYU University of Physics and Mathematical Sciences, Mar. 12, 2009.

[4] Shunzhi Zhu, Yan Wang, Yun Wu, "Health Care Fraud Detection Using Nonnegative Matrix Factorization", The 6th International Conference on Computer Science Education (ICCSE 2011) August 3-5, 2011. SuperStar Virgo, Singapore.

[5] Zhongyuan Zhang, Tao Li, Chris Ding, Xiangsun Zhang, "Binary Matrix Factorization with Applications", Proceeding ICDM '07 Proceedings of the 2007 Seventh IEEE International Conference on Data Mining Pages 391-400.

[6] Mohammad SajjadGhaemi. Class Lecture, Topic: "Clustering and Nonnegative Matrix Factorization". DAMAS LAB, Computer Science and Software Engineering Department, Laval University. Apr.12, 2013.

[7] Haesun Park. Class Lecture, Topic: "Nonnegative Matrix Factorization for Clustering". School of

Computational Science and Engineering Georgia Institute of Technology Atlanta, GA, USA, July 2012.

[8] Fashoto Stephen G., OwolabiOlumide, Sadiku J., Gbadeyan Jacob A, "Application of Data Mining Technique for Fraud Detection in HealthInsurance Scheme Using Knee-Point K-Means Algorithm", Australian Journal of Basic and Applied Sciences, 7(8): 140-144, 2013 ISSN 19918178.

[9] Leonard WafulaWakoli. "APPLICATION OF THE K-MEANS CLUS-TERING ALGORITHM IN MEDICAL CLAIMS FRAUD/ABUSE DETECTION," MSc Thesis, Jomo Kenyatta University Of Agriculture And Technology, 2012.

[10] Guido Cornelis van Capelleveen, "Outlier based Predictors for Health Insurance Fraud Detection within U.S. Medicaid", University of Twente University of California, San Diego December 2013.

[11] Qun Song, Nikola Kasabov, "ECM A Novel On-line, Evolving Clustering Method and Its Applications", Department of Information Science, University of Otago. H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed.Harlow, England: Addison-Wesley, 1999.