ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Survey on Hybrid Approach for Fraud Detection in Health Insurance

Punam Devidas Bagul, Sachin Bojewar, Ankit Sanghavi

M. E Student, Dept. of Computer Science and Engineering, ARMIET, Sapgaon, Mumbai University, India.

Professor, Dept. of Information Technology, VIT, Wadala, Mumbai University, India.

Head of Department, Dept of Computer Science and Engineering, ARMIET, Sapgaon, Mumbai University, India.

ABSTRACT: Any act committed with the intent to obtain a fraudulent outcome from an insurance process is referred as Insurance fraud. When a claimant attempts to obtain some benefit or advantage to which they are not entitled then that attempt is considered as insurance fraud. Since last decades, Fraud and abuse on medical claims became a major concern for health insurance companies. Data mining techniques are applied to detect and avoid such frauds. This includes analysis of the characteristics of health care insurance data, some preliminary knowledge of health care system and its fraudulent behaviors. This paper surveys various Data mining techniques applied to various applications. Data mining technique is divided into supervised and unsupervised learning techniques to detect fraudulent claims. But, each of the above techniques has its own set of advantages and disadvantages, So a novel hybrid approach for detecting fraudulent claims in health insurance industry is need to be proposed, by combining the advantages of both the techniques.

KEYWORDS: Health insurance fraud, Data mining, Supervised, Unsupervised, Hybrid approach.

I. Introduction

Health insurance fraud is an intentional act of deceiving, concealing or misrepresenting information that makes benefit to an individual or group. Financial benefit is the main purpose of fraud. It is estimated that the number of false claims is approximately 15 per cent of total claims as per the recent survey. Hence health insurance fraud detection becomes challenging task. So, it is necessary to minimize or eliminate fake claims to make health insurance industry free from fraud.

Health insurance fraud claims can be

- Billing services that are costlier than the actual procedure that performed.
- Billing for medical equipment that is costlier than the actual equipment.
- Billing for services that are not covered under policy coverage.

To detect these frauds Data mining techniques are applied. Data mining technique is divided into supervised and unsupervised learning techniques to detect fraudulent claims. But, each of the above techniques has its own set of advantages and disadvantages. This paper surveys advantages, disadvantages of various data mining techniques. So finally it is need to develop novel hybrid approach for detecting fraudulent claims in health insurance industry by combining the advantages of both the techniques.

DOI: 10.15680/IJIRCCE.2016. 0404092

II. MOTIVATION

Fraud is costly and widely spread in Health Insurance Company. Traditionally, health insurance companies used heuristic rules to detect frauds. These rules were summarized from previous fraud cases and are used to detect fraud either through human inspection or interaction with an external entity. Traditional fraud detection approaches gets fail because of increasing sizes of databases. It is shocking that the incidence of health insurance fraud keeps increasing every year. So it is need to propose a new advance approach to detect suspicious health care frauds from large databases.

III. LITERATURE SURVEY

This section provides the study of Data mining approaches and their respective advantages, disadvantages.

Prasad Seemakurthi, Shuhao Zhang and Yibing Qi Performed work on "Detection of Fraudulent Financial Reports with Machine Learning Techniques" [1]. They give an overview of advanced supervised machine learning and natural language processing techniques, including Binomial Logistic Regression, Support Vector Machines, Neural Networks, EnsembleTechniques, and Latent Dirichlet Allocation (LDA), to the problem of detecting fraud in financial reporting documents.LDA is a generative probabilistic model whose basic idea is that documents are represented as a random mixture over latent topics, and each topic is characterized by a distribution of words. LDA model extracted the document-topic distribution matrix which contains the Dirichlet probability of each document.

To handle such nonlinear data, SVM and neural network are applied. Logistic regression is conducted to compare with two classifiers.

Finally it implemented ensemble technique whose input is the output of each of algorithm.

The authors in [2] detect suspicious health care frauds from large databases using clustering technique. It applies two clustering methods SAS EM and CLUTO to health insurance dataset and compares their performances.

CLUTO is able to cluster various datasets in diverse application areas and it handles high-dimensional datasets very well.

SAS EM creates accurate predictive and descriptive models that are based on analysis of large amounts of data from the enterprise.

As perthe experimental results CLUTO is faster than SAS EM and SAS EM gives more useful clusters than CLUTO.

[3] "Fraud Detection in Health Insurance using Data Mining Techniques" states that SVM (Support Vector Machine) and Evolving Clustering Method (ECM) are applied in health insurance field for fraud detection.

In this paper Support Vector Machine is algorithm used for classification and Evolving Clustering Method algorithm used for clustering.

SVM technique trained the system to determine decision boundary between "legitimate" and "fraudulent" claims classes whileIn ECM**,** as and when new data point comes in, ECM clusters them by modifying the position and size of the cluster.

SVM provides high accuracy and work well though data is not linearly separable but it has high complexity.ECM is used to cluster dynamic data hence it find out newly incoming fraudulent claims.

Hossein, Arash, Mahmood, Mahdi, 2015 [4] In this paper data mining techniques are performed for detecting health care fraud and abuse with the help of supervised and unsupervised data mining techniques.

The supervised methods applied to health care fraud and abuse detection are decision tree, neural networks, genetic algorithms and Support Vector Machine (SVM). The unsupervised methods that have been applied to health care fraud and abuse are clustering, outlier detection and association rules.It recommends outlier detection as unsupervised method and routine online processing task as supervised learning method.

Authors in International conference on August 3-5 2011, Singapore [5] introduces a technique for clustering semantic features in a prescription collection and form clusters of medical treatment items on the basis of shared semantic features to detect fraud. This technique is referred as Non-negative Matrix Factorization (NMF) method.

DOI: 10.15680/IJIRCCE.2016. 0404092

The factorization is preserving natural data non-negativity as well as to compute a low rank approximation of a large sparse matrix.

NMF algorithm is used to cluster medical items according to monthly cost on them.

Reference [6] paper gives effective method for health insurance fraud detection that identifies suspicious behaviour of health care providers is outlier based predictor.

Outlier detection methods used were deviation from regression model, trend deviations, peak deviations, deviation clusters and single deviations from clusters.

This technique detected the frauds with 71% accuracy.

"Application of Bayesian Methods in Detection of Healthcare Fraud" [7] Application of Bayesian ideas are presented in healthcare fraud detection. Bayesian co-clustering is used to identify fraudulent providers and beneficiaries who have unusual behaviour.

This paper proposed that detection of this type of unusual behaviour will be helpful to decision makers in audits.The Bayesian technique provides formalism for both making decisions for investigation of fraud as well as quantifying uncertainty about fraudulent behaviour. This also reduces cost of investigation.

The use of Bayesian technique would be help in finding future evolution of clusters and forecasting the future behaviour of new providers or beneficiaries as per their given characteristics.

Because of the difficulty in accessing medical data due to confidentiality and privacy issues, it is challenge to use such statistical approaches in healthcare fraud detection system. Finally this paper stated that advancement in statistical approaches in medical fraud assessment which can combine with medical prevention, detection and response needed.

The paper in [8] compared the efficiency between number of classification methods such as Decision Tree (DT), Logistic Regression (LogR), k-Nearest Neighbor (k-NN), Naïve Bayes (NB), C4.5) and Linear Classifier (LC) with respect to the Area Under Curve (AUC) metric.Parameters for comparison were size of the dataset, type of the independent attributes, and the total number of the continuous and discrete attributes. Authors in [8] concluded that datasets containing less number of records then AUC becomes deviate so comparison not possible. The comparison was possible only when the number of the records and the number of the attributes in each record were increased.

Authors in [8] gave result of comparison that C4.5 gives higher AUC in the most cases. It is said that NB provides the best AUC among LogR and LC. Classifier LogR and Classifier NB have the same results, approximately.

The future scope of this paper was to compare the efficiency of other classifiers by using the current method.

Hetal, Amit in [9] worked on research that is related to the study of the existing classification algorithm and their comparative study in terms of accuracy, speed, scalability and other issues. This study helps in studying the existing algorithms as well as developing new innovative algorithms for new application.

Comparative study of very well-known classification algorithms involved Decision Tree Induction, Neural Network, K-nearest neighbours Bayesian Network, and Support Vector Machine. This comparative study focused on advantages and disadvantages of one method over other method and their own area of implementation. No one algorithm satisfied all the required criteria.

So this paper concluded that new classifier is needed to investigate which can be built by combining two or more classifier's strength.

IV. PROPOSED SYSTEM

Many fraud detection systems used either supervised or unsupervised learning algorithms. Implemented Supervised learning algorithms are Support Vector Machine, Neural networks, Logistic regression. But no single algorithm among them gave satisfactory results. Supervised learning algorithms are failed to detect unexceptional conditions which can be handled if unsupervised learning algorithm is conducted.
While some fraud detection system are implemented only unsupervised learning algorithms like Outlier detection, Evolving Clustering method. But because of lack of direction, sometimes no interesting knowledge can be discovered with only unsupervised algorithm.

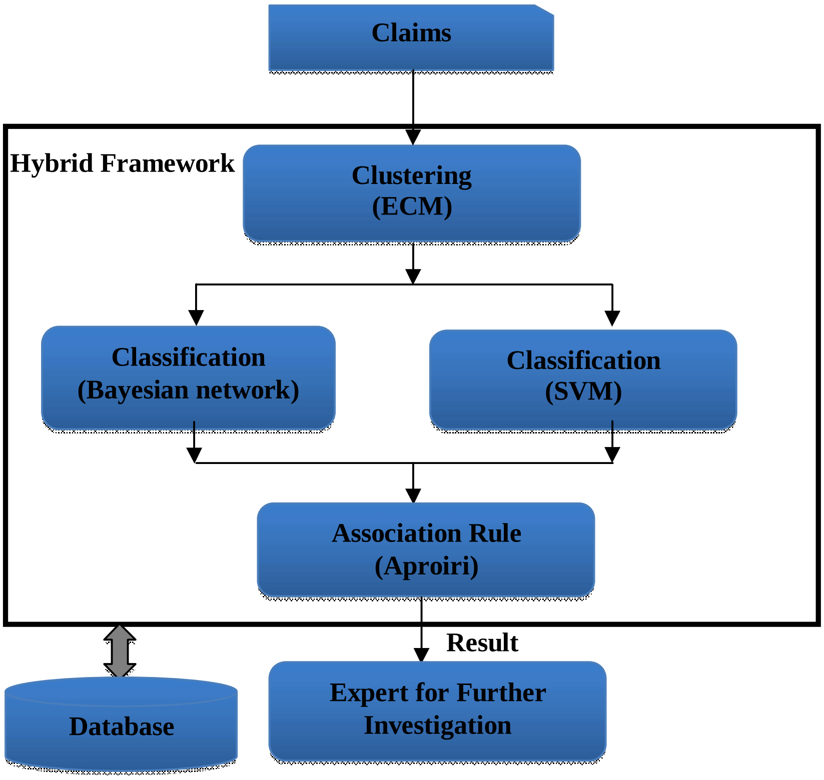
Claims
Hybrid Framework
Clustering (ECM)
Classification (Bayesian network)
Classification (SVM)
Association Rule (Aproiri)
Result
Database
Expert for Further Investigation

Fig. 1. Proposed Architecture

So I proposed a hybrid model shown in figure 1 that uses combination of both supervised and unsupervised learning algorithms to detect fraudulent claims and makes it even more efficient with association rule mining. This Proposed Model will try to give the best results as far as possible and will satisfy customer's requirement.

V. Conclusion and Future Work

Data mining is a technique that extracts knowledge from large databases of thousands of claims. It identifies a smaller subset of the claims or claimants for further assessment fraud detection. In this way, the data mining approach is most appropriate technique for fraud detection in health insurance industry. Data mining techniques are categorised into supervised and unsupervised techniques. As per the research survey it is cleared that no any single technique is sufficient for our application because each technique having its own set of advantages and disadvantages. So combining advantages of both supervised and unsupervised technique definitely makes system better.

VI. ACKNOWLEDGEMENT

I owe a great many thanks to all who helped and supported me during development of this paper. I sincerely thank my project guide Prof. Sachin Bojewar sir for his expert guidance and valuable support throughout the course of this project work. I also acknowledge with deep sense of gratitude and inspiration received from my head of computer department and principle of my college.

References

1. Prasad Seemakurthi, Shuhao Zhang, and Yibing Qi,"Detection of Fraudulent Financial Reports with Machine Learning Techniques", IEEE Systems and Information Engineering Design Symposium University of Virginia, 2015.

2. Yi Peng, Gang Kou, Alan Sabatka, Zhengxin Chen, Deepak Khazanchil, Yong Shi,“Application of Clustering Methods to Health Insurance Fraud Detection” , IEEE paper.
3. Vipula Rawte, G Anuradha, “Fraud Detection in Health Insurance using DataMining Techniques”, International Conference on Communication, Information & Computing Technology (ICCICT), Jan. 16-17,2015.
4. Hossein Joudaki, Arash Rashidian, Behrouz Minaei-Bidgoli, Mahmood Mahmoodi, Bijan Geraili, MahdiNasiri,“Using Data Mining to Detect Health Care Fraud and Abuse”,Global Journal of Health Science Vol. 7, No. 1; 2015
5. Shunzhi Zhu, Yan Wang, Yun Wu, **“**Health Care Fraud Detection Using Nonnegative Matrix Factorization”, International conference on August 3-5 2011, Singapore.
6. Guido Cornelis van Capelleveen**, “**Outlier based Predictors for Health Insurance Fraud Detection within U.S. Medicaid”, at the University of Twente December 2013.
7. Tahir Ekina, Francesca Levab, FabrizioRuggeri c, Refik Soyer d, “Application of Bayesian Methods in Detection of Healthcare Fraud” Chemical Engineering Transactions Vol. 33, 2013.
8. Reza Entezari-Maleki, Arash Rezaei, and Behrouz Minaei-Bidgoli, “Comparison of Classification Methods Based on the Type of Attributes and Sample Size”.
9. Hetal Bhavsar, Amit Ganatra, “A Comparative Study of Training Algorithms for Supervised Machine Learning”International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, September.