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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Generating Optimal Result for Fraud Detection Using Data Mining

Punam Bagul, Sachin Bojewar

M. E Student, Department of Computer Engg, ARMIET, Mumbai University, India

Associate Professor, Department of Information Technology, VIT, Mumbai University, India

ABSTRACT: Data mining is widely used in health insurance companies for extraction of implicit, previously unknown and potentially useful information from data. The system uses clustering and classification approach for analyzing characteristics of health care insurance data, some preliminary knowledge of health care system and its fraudulent behaviors. This data mining approach helps a user to detect fraud claims. In this I am proposing an approach in which user enters required details from new claim form and based on claim form details, clustering approach forms clusters of claims having similar characteristics. Classification technique trained the system to determine a decision boundary between classes of legitimate and fraudulent claims. Then each claim details are compared with that decision boundary and is placed into either legitimate or fraudulent class.

KEYWORDS: Data mining, Clustering, Classification, Hybrid approach, SVM, Evolving Clustering Method and Bayesian.

I. INTRODUCTION

In order to obtain unauthorized benefits, an intentional deception used is referred as health care fraud. Data mining is a technique that helps health insurance organizations to extract useful knowledge from raw thousands of claims. It helps to identify a smaller subset of the claims for further assessment and for detection of fraud and abuse.

Data mining technique is basically divided into two machine learning techniques viz., supervised and unsupervised is implemented to detect fraudulent claims. But, since each of the above machine learning techniques has its own set of advantages and disadvantages. So there was need of proposing a novel hybrid approach for detecting fraudulent claims in Health Insurance Company by combining the advantages of both the learning techniques.

To implement fraud detection system, we trained the system using some previously known claim data details. This previously known claim data is referred as training dataset. System performs clustering process on training dataset. Then classification technique is applied on results of clustering. System compares each new incoming claim with unknown result with trained dataset and gives result accordingly.

This paper presents an approach to generate optimal result for fraud detection in Health Insurance Company. In section 2, the literature survey of the past work on fraud detection is presented. Proposed system is discussed in section 3. And paper is concluded in section 4.

II. RELATED WORK

This section includes the relevant past literature that uses the various data mining approaches for fraud detection in health insurance. Most of the approaches trained system with machine learning data mining approaches.

Authors Yi Peng, Gang Kou, Alan Sabatka, Zhengxin Chen, Deepak Khazanchil, Yong Shi [1] proposed a mechanism to find out suspicious health care records using clustering unsupervised technique. SAS EM and CLUTO is used for clustering large databases. CLUTO handled high dimensional database easily and also applicable to various datasets in diverse application areas. SAS EM creates accurate predictive analysis of large amounts of data from the enterprise

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Hossein Joudaki, Arash Rashidian,Behrouz Minaei-Bidgoli, Bijan Geraili, Mahdi Nasiri & Mohammad Arab [2]focused on the approach which will used data mining approaches that includes both supervised and unsupervised machine learning methods. Supervised methods used were decision tree, neural networks, genetic algorithms and Support Vector Machine (SVM). Clustering, outlier detection and association rules were used as unsupervised learning technique for fraud and abuse detection in healthcare. This paper recommends combination of both machine learning techniques.

The approach proposed in this project uses both machine learning methods i.e. supervised as well as unsupervised.

As in [3] Vipula Rawte, G Anuradha have proposed that In order to develop fraud detection system with efficient performance has to use the claim details from claimant's past behaviors so that it can predict new claim's category. This paper introduced an approach in which system will learn rules from training dataset using classification based on previously known labels. SVM (Support Vector Machine) is fundamentally a classification technique recommend here. It suggested ECM clustering technique when there is a need to cluster dynamic data and SVM classification technique is used for its scalability and usability to detect fraudulent claims.

Guido Cornelis van Capelleveen [4] gives effective method for health insurance fraud detection that identifies suspicious behavior of health care providers is outlier based predictor.

III. PROPOSED SYSTEM

We proposed model that applies clustering technique (Evolving Clustering Method ECM) to form clusters depends on the disease type for each of the newly incoming health insurance claim. Clustering is followed by two classification techniques viz. Support Vector Machine (SVM) and Bayesian network. SVM classify those claims into either legitimate or fraudulent class. Bayesian network encodes probabilistic relationship on that data.

Dataset consists of the data collected at hospital level. Once system is trained by training dataset, new claim is then classified using SVM and gives result to user.

A. *The proposed approach consists of three phases:*

1. Apply Clustering technique to form clusters.
2. Apply Bayesian network for Classification of clusters.
3. Applying Support Vector Machine for further classification.

B. *Extraction of claim details:*

This phase analyses the claim details given by user in the form of claim form filled by claimant. Claimant form includes disease, disease type, medication, medicine type, claimant amount. These types of details are required to collect similar type of results from training dataset.

Data Extraction:

Dataset consists of claim details submitted on claim form is collected and added in the database. It includes details that are more susceptible to determine fraud. Data extraction displays dataset to user.

C. *Clustering Technique to form Clusters:*

This module accepts the training dataset details and converted into numerical values. Then each dataset has its own numerical values. Based on these values Euclidian distance is calculated. Various clusters are formed depends on Euclidian distance. Euclidian distance between two points A(x1, y1), B(a, b) in space is given by

$$\text{Euclidian distance} = \text{sqrt}((x1-a)^2 + (y1-b)^2).$$

D. *Applying Bayesian network* :

Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. It calculates the mean and variance for course duration, hospitalization days and medicine cost.

With result of mean, variance it calculates the standard deviation which determine probability distribution.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

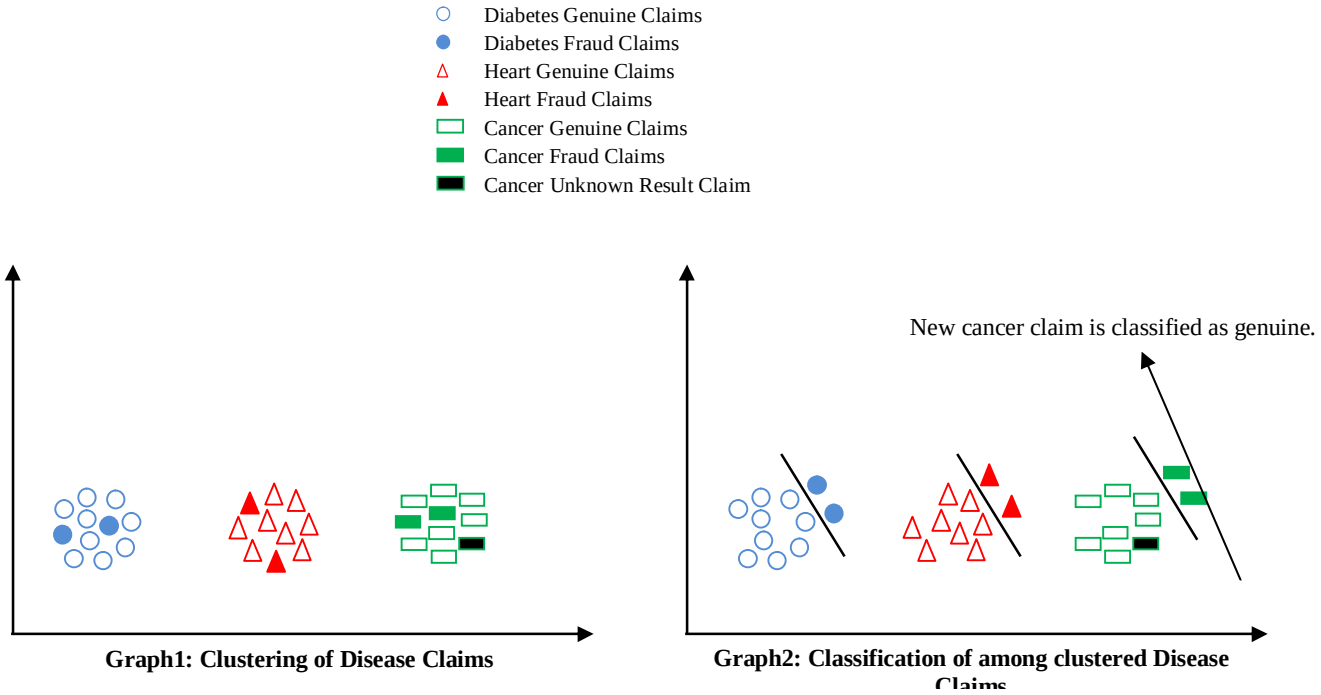
Vol. 5, Issue 5, May 2017

E. *Applying Support Vector Machine:*

It is another classification method used. It classifies both linear and nonlinear data. To classify nonlinear data, nonlinear mapping function is used. It transforms the original training data into a higher dimension data. With the new higher dimensional data, it searches for the linear optimal separating hyperplane (i.e., "decision boundary"). With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane.SVM finds this hyperplane using support vectors ("essential" training tuples) and margins (defined by the support vectors.

IV. RESULT

Dataset contains all claim records whose results are previously known is submitted to train the system. Clustering is applied on the dataset and new claim submitted of unknown result. Classification is then followed by clustering. Classification uses combined result of both classification techniques i.e. Bayesian network and SVM.



Graph1: Clustering of Disease Claims

Graph2: Classification of among clustered Disease Claims

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed fraud detection system which learns from the past claimant's claim details. The goal of the system is to provide optimal results. Furthermore, an attempt is made to apply two classification algorithms along with the clustering algorithm. As the data set considered for this implementation is health insurance dataset. In future this system can be extended to other domain where user needs a good fraud detection system.

VI. ACKNOWLEDGMENTS

Our sincere thank to the experts who have contributed towards development of this paper. I remain immensely thankful to Prof. Sachin Bojewar for his invaluable support in garnering resources for me either by way of information, for providing me with the idea of this topic, and or computers also guidance and supervision.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

REFERENCES

1. Yi Peng, Gang Kou, Alan Sabatka, Zhengxin Chen, Deepak Khazanchil, Yong Shi "Application of Clustering Methods to Health Insurance Fraud Detection".
2. Hossein Joudaki, Arash Rashidian,Behrouz Minaei-Bidgoli, Bijan Geraili, Mahdi Nasiri & Mohammad Arab(2015) "Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature" Global Journal of Health Science; Vol. 7, No. 1; 2015
3. Vipula Rawte, G Anuradha "Fraud Detection in Health Insurance using Data Mining Techniques" 2015 International Conference on Communication, Information & Computing Technology (ICCICT), Jan. 16-17, Mumbai, India
4. Guido Cornelis van Capelleveen "Outlier based Predictors for Health Insurance Fraud Detection within U.S. Medicaid".
5. Shunzhi Zhu Yan Wang, Yun Wu (2011) "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.
6. Guido Cornelis van Capelleveen "Outlier based Predictors for Health Insurance Fraud Detection within U.S. Medicaid".
7. Reza Entezari-Maleki, Arash Rezaei, and Behrouz Minaei-Bidgoli "Comparison of Classification Methods Based on the Type of Attributes and Sample Size".
8. Jiawei Han, Micheline Kamber, "Data Mining Concepts and techniques".