IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A TYPICAL STUDY OF IMPROVING ACCURACY IN DETECTING INSURANCE FRAUD ON UNSTRUCTURED DATA SETS

N.Pratheeba*, N.Dhanalakshmi

Department of Computer Science and Engineering, PSNA College of Engineering and Technology,
Dindigul, Tamilnadu-624619, India.

ABSTRACT

Fraud in insurance health care brings significant financial and personal loss on individuals, business, government and society as a whole. The size of health care sector and the enormous volume of money involved make it an important fraud target. The big data trend, (the growth in unstructured data) always leaves lots of rooms for a fraud going undetected if data is not analyzed properly. Performing big data analysis can identify repetitive errors that are hidden and prevent the occurrence of them in future. The primary objective of this paper is to define existing challenges of fraud detection for the different types of large data sets and ways to extract the features that cause fraud. It also deals with the methods for improving accuracy by considering both true positives and true negatives, thereby performing data analytics.

Keywords: Insurance fraud, Health care, Classification, Spatial hypothesis, Accuracy, Data analytics.

INTRODUCTION

Data mining procedures are utilized as a part of a numerous application ranges, artificial intelligence, hereditary qualities and advertising. The information mining procedure comprises of three levels: (1) The beginning investigation, (2) model building or example distinguishing proof with confirmation/approval and, (3) Arrangement (i.e., predictions generated by the utilization of the model to the new data). On the other hand, a typical general distinction in the center and reason between Information Mining and the Exploratory Information Investigation (EDA) is that Information mining is more concerned towards applications than the fundamental way of the basic wonders.

As it were, information mining has less worry about distinguishing the particular relations among the included variables. Insurance is an understanding in which an individual or a gathering gets security against misfortunes from an insurance agency. Insurance agencies are utilizing data analytics for fraud discovery. Henceforth, harnessing digitization is a critical variable. Handling of fraud manually has always been costly for insurance companies, even when few low incidences of high-value fraud were undetected. While developing machine learning calculations to binary data problems, data unevenness has become a challenge to investigators. The system for sorting out machine learning algorithms is valuable that it makes us to consider the parts of the information and the model arrangement handling and select one that is the most proper for your issue keeping in mind the end goal to get the best result. Supervised or administered learning comprises of a dependent variable (or target variable) which is to be predicted from a given arrangement of independent (free variables). Utilizing these variables, we create a capacity to outline inputs to the desired results.

The training procedure proceeds until the model accomplishes a sought level of precision on the trained data. Samples of Supervised Learning are Decision Tree, Regression, Random Forest, KNN, Logistic Regression etc. While, unsupervised learning don't have any objective variable to anticipate or predict. It is utilized for clustering samples for diverse groups, which is generally used for partitioning similar groups based on their similarity measures. Illustrations of Unsupervised Learning are Apriori algorithm and K-means. In the ongoing scenario, the volume of information utilized straightly increments with time. By organizing the information it has the capacity recognize mistaken or suspicious records in submitted social insurance information sets and gives a methodology of how the doctor's facility and other human services information is useful for the implementing so as to distinguish medicinal services insurance fraud, for example, decision tree, clustering and naive Bayesian classification. Thus the survey that will determine the accuracy by utilizing spatial theory relationship of deceitful records and removing the qualities that causes misrepresentation is presented in this paper.

The remaining of this paper is organized as follows: Section 2 reviews brief about the existing methods for detecting insurance fraud in various sectors, the techniques used with pros and cons. Section 3 presents the challenges of unstructured data and ways to structure them. Section 4 details the validation process by removing outliers. Section 5 presents spatial analysis and feature extraction for data analytics. Finally, Section 6 concludes this paper.

LITERATURE REVIEW

In this segment, we audit the works related to fraud detection. A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance uses krnn (k reverse nearest neighborhood)to recognize and to eliminate the noise present as anomalies in the dominant part class, then ocsvm (one class support vector machine)to concentrate support vectors in the larger part class as shown in Fig. 1. At last, combined these redundant and noise free larger part class samples with the original minority samples and performed experiments with the modified dataset therefore got. In this way it removes anomalies (outliers) and repetitive examples from the dominant part class of profoundly skewed unbalanced datasets. Likewise, Acquires high concentration on sensitivity by utilizing few number of standards. Yet, it experiences the downsides of specificity since sensitivity is concurred top need in front of specificity[1].

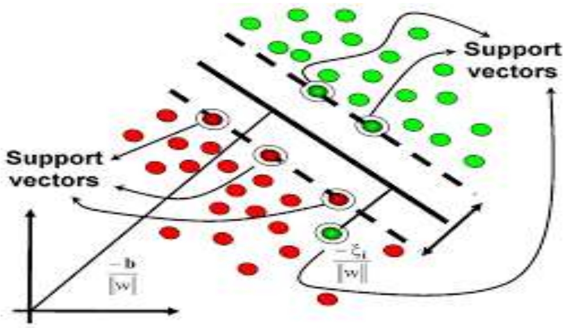

Figure 1. One class support vector machine.

Topological pattern discovery and feature extraction for fraudulent financial reporting gives an effective classification standard to recognize FFR in light of the topological patterns and a specialist aggressive element extraction component to catch the striking attributes of fraud practices. This depends on Growing Hierarchical Self-Organizing Map (GHSOM), an extension of Self-Organizing Map (SOM) which is an unsupervised neural system for grouping. Be that as it may, the model is connected just for trained topological samples. The determined rules are dependent on data and can't be straightforwardly connected to other data contexts that neglect to fulfill the spatial hypothesis [2].

Using SOM and PCA for analyzing and interpreting data utilizes two-dimensional reduction methods, Kohonen SOM and PCA for the multivariate examination, analysis and procedure understanding of an information set and altogether analyze and interpret multidimensional data as shown in Fig. 2. It has a clear understanding and capacity to manage nonlinear issues. The technique is very effective to suppress multidimensional data sets. Likewise gives good representation capabilities and highlights the likeness among the observations. No control algorithm is produced from the separated data keeping in mind the end goal to advance procedure execution and not sensing the online estimations from the cheap and low maintained sensors to computerize process operation by the control procedure [3].

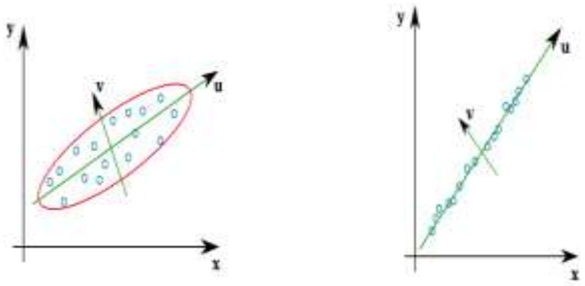

Figure 2. PCA for data representation and dimensionality reduction.

A new stability concept to detect changes in unsupervised data streams.This depends on the surrogate data methodology from time arrangement examination gives training to online unsupervised calculations even if there should be an occurrence of time reliance among observations. Having the capacity to figure out if or not information attributes are changing along time is a noteworthy towards data stream algorithms. Algorithm 1. Recognizing whether two time arrangement offer comparative Qualities, Algorithm 2. Stable change recognition calculation. It effectively figures out if or not data qualities are changing along time arrangement furthermore identify changes in unsupervised data streams. Yet, it makes no investigation of distinctive separation capacities to figure the dissimilarity between PS diagrams, what may enhance results for a few applications in which recurrence varieties are adequate [4].

An Analytical Approach To Detect Insurance Fraud Using Logistic Regression on accident coverage fraud, which happens in both auto physical damage (APD-collision and comprehensive) and injury claims (Personal Injury Protection-PIP)looking at different circumstances inside APD and PIP claims and different strategies that safeguarded individuals use to cheat insurance agencies[5]. Here Cases are organized in view of the rate probabilities as shown in Fig. 3. that show higher chances of fraud.

Better prioritization of cases that are to be researched. The logistic model in foreseeing fraud is just an instrument and that the determination of fraud would not be based on the model but rather when complete and exhaustive examination.

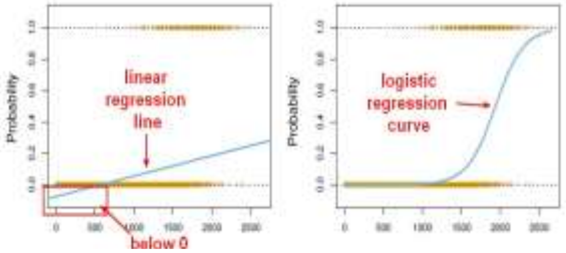

Figure 3. Linear regression output as probabilities.

Healthcare Insurance Fraud Detection Leveraging Big Data Analytics, a way to detect and foresee potential fraud issues by applying bid data, hadoop environment and diagnostic methods which can prompt fast identification of fraud anomalies. The result depends on a high volume of historic data from different insurance agencies and clinic information of a particular geological area [6]. Analytic modules, for example as decision tree, clustering and naive Bayesian classification are utilized. The technique has the capacity to identify wrong or suspicious records in submitted medicinal services data sets and gives a methodology of how the doctor's facility and other social insurance information are useful for recognizing human services insurance fraud. The procedure additionally minimizes the loss of assets to fraud cases. The proposed system neglected to get precision in detecting fraud and not intended to be sufficiently alterable to adjust the changes.

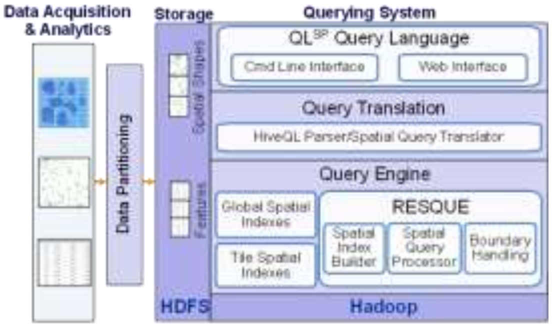

Figure 4.Architecture of Hadoop.

An expert system for detecting automobile insurance fraud using social network analysis presented systems are represented in form of networks such as social area, permitting definition and examination of complex relations between entities. Fraudulent details are found by utilizing a novel appraisal, Iterative Assessment Algorithm (IAA). The framework permits the training of experts domain knowledge even without named data set and it can accordingly be adopted to new sorts of fraud cases when they are taken note. The cons of this framework are no certification that the limit n/2 is the best decision. It doesn't consider what number of segments has some specific indicator set and all indicators are dealt with equal importance [7].

Auto Claim Fraud Detection Using Multi Classifier System utilizes a cost matrix and a mix of classifiers. This work recognizes the most cost saving model to perform the identification of associated cases with fraud in a dataset of vehicles cases. The framework uses Cost Grid mix of classifiers (C4.5, SVM and Naive Bayes calculation) to anticipate the last order of the every object in the dataset. This work joined the consequence of every algorithm beforehand introduced to distinguish associated cases with fraud behaviors by parallel topology. This is a temperate model to perform the discovery of suspected cases with fraud [8]. Utilizing blend of classifiers as a part of a parallel topology makes this system more effective. In any case, proficiency is not enhanced since it utilizes imbalanced classes of data sets.

Anomaly Detection via Online Oversampling Principal Component Analysis by oversampling the objective occurrence and removing the foremost direction of the data, the proposed osPCA (Online oversampling principal component analysis) permits us to decide the abnormality of the objective instance as per the variety of the subsequent dominant eigenvector. Since osPCA need not perform Eigen investigation expressly, the proposed system is favored for online applications which have calculation or memory confinements. Online oversampling principal component analysis (osPCA) algorithm[9] is utilized to identify the deviations of exceptions from a lot of data by means of an online upgrading system. It decides the inconsistency of the target sample as per the variety of the subsequent dominant Eigen vector without storing the entire data matrix or covariance.Thereby, Lessens computational expenses and memory. However the framework is not best in evaluating the main principal directions to handle data in multi dimensional space.

Unstructured Data Analysis on Big Data using Map Reduce manages the huge volume of data; deals with the colossal volume of information. The proposed system will prepare the data in parallel as small chunks and total all the information across the clusters in obtaining processed data.

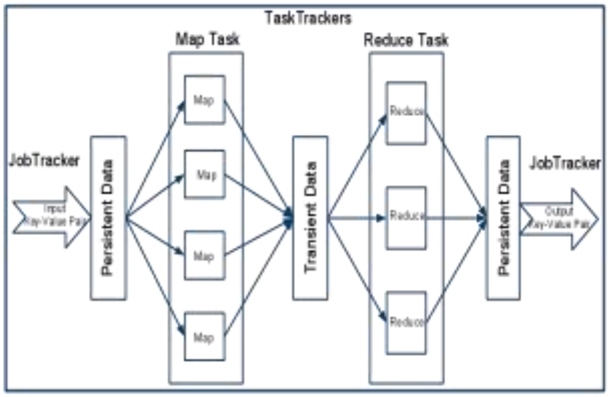

Figure 5.Map Reduce framework.

In Hadoop system, map reduce is utilized to perform the task of aggregation and filtering and to keep up the proficient storage structure. The data are ideally refined utilizing collaborative filtering, under the prediction of specific information required by the client. The proposed system is upgraded by utilizing the procedures, for example, sentiment analysis through common language processing for parsing the information into tokens and emoticon based grouping. The procedure of data grouping depends on client feelings to get the information required by a particular client. The unstructured data is organized and handled. The map reduce strategy utilized conquers the capacity issues in light of the fact that dissimilar to hash calculation it needn't bother with a great deal of space for putting away the hash values and hash tables. The framework has the disadvantage of not executing map reduce work in distributed mode in which we can utilize a N number of slaves for a single master[10].

DATA STRUCTURING

Unstructured data, represented in binary form cannot be identified by its internal structure. Many people spend half of their days in preparing powerpoint,spreadsheets,email and other unstructured data. Moreover, structured database not even contain half of the information useful to people. Therefore the right data doesn't reach people in a correct time. Only the valuable time spent on it is wasted. Unstructured data has more missing values, hence missing the value of those data is not that much easy since data can't be analysed, searched, sorted and visualized. These types of data need extra processing time and new tools to extract the information and to deliver it. Structured data, the opposite of unstructured can be fit into a database. Hence it is also termed as relational data. These types of data can be mapped easily to their relevant fields.

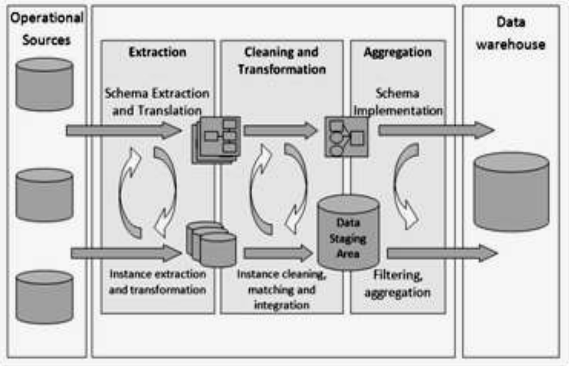
Operational Sources
Extraction
Cleaning and Transformation
Aggregation
Data warehouse
Schema Extraction and Translation
Schema Implementation
Data Staging Area
Instance extraction and transformation
Instance cleaning, matching and integration
Filtering, aggregation

Figure 6. Data Preprocessing.

Hence there is a need to transform unstructured data to structured data using data mining techniques such as data preprocessing, data cleaning, data reduction, data transformation, and finding missing values. Over past few years they have recognized that structured data has some value. Health information data uses standards for data interpretation and converting it to structured data.

OUTLIER ANALYSIS

Outlier is an abnormal deviation from other set of values in a random population. Machine learning algorithms are sensitive to the range and distribution of attribute values in the input data. Outliers in data set mislead training process of supervised learning resulting in poorer results with less accuracy and longer training time. Few tests of outlier are to detect the presence of single outlier while others test the presence of multiple outliers. It is not an appropriate method to apply a test for a single outlier sequentially to detect multiple outliers.

Health insurance fraud is quite prevalent, because of the ease with which sensitive information number may be compromised. This typically leads to unauthorized use of the insurance policies. In many cases, unauthorized use may show different patterns. Such patterns can be used to detect outliers in healthcare insurance data. Many medical applications collect the data from a variety of devices such as MRI scans, PET scans or ECG time-series. Unusual patterns in such data typically reflect disease conditions. The process of identifying outliers has many names in data mining and machine learning such as outlier mining, outlier modeling and novelty detection and anomaly detection.

Some commonly used methods for outlier detection are as follows:

Probabilistic and Statistical Models: Determine the unlikely instances from a probabilistic model of the data. For example, Gaussian mixture models optimized using expectation-maximization.

Linear Models: Models the data into lower dimensions using linear correlations. For example, large residual errors may be outliers with principal component analysis and data.

Extreme Value Analysis: Determine the statistical extremes of the underlying distribution of the data. For example, like
Z-scores on univariate data using statistical methods.

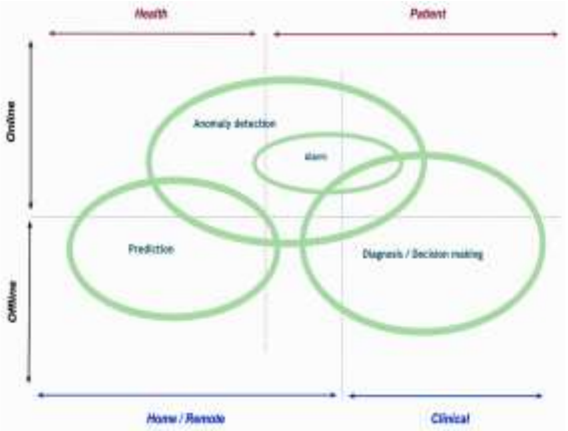

Figure7.Anamoly detection in health monitoring.

The abnormal patterns are significantly valuable in the insurance domain. Detected unusual patterns in health parameters allow making accurate decisions in short time. Anamoly detection methods are based on classification techniques to differentiate the data set into normal class and outliers. Support vector machines, Markov models and wavelet analysis are used in detecting fraud.

V.SPATIAL HYPOTHESIS AND FEATURE EXTRACTION

The ultimate goal in detecting fraud is to provide evidence based insight through deeper understanding of data and to produce results that can be used at policy and strategy levels. Spatial data mining same as data mining but with the end objective to find patterns. Next, by applying and extending these techniques helps us to solve difficult image feature extraction problems.

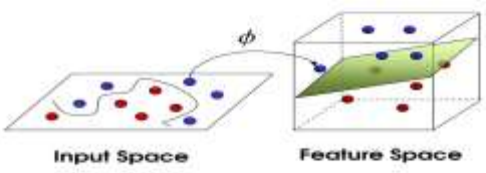

Figure 8. Feature extraction.

By performing spatial hypothesis, dichotomous samples (fraud and non-fraud) can be grouped into separate clusters. While performing clustering the characteristics or features that cause fraud also can be identified. The extracted features can be used to prevent the occurrence of the fraud samples in insurance data to happen in future by performing data analytics.

CONCLUSION

Data analytics and balancing the data brings fraud detection in insurance to another level. Earlier, investigations on insurance fraud were too costly and took more time. So many companies prefer to pay claims without investigation.

This paper presented the strength and weakness of the existing fraud detection techniques. The analysis methods applied in field oh health insurance were briefly described, each of them being effective for a particular type of fraud or a particular stage of fraud detection process. The survey presented in this paper will help to proceed our future work on detecting insurance fraud in real work healthcare application.

REFERENCES

[1] Aguado , Montoyaa, Borras, Secob and J. Ferrer(2007),’Using SOM and PCA for analysing and interpreting data from a P-removal SBR’,ELSEIVER.

[2] Ganesh Sundarkumar and Vadlamani Ravi(2014),’A novel hybrid undersamplingmethod for mining unbalanced datasets in banking and insurance’,ELSEIVER.

[3] Holton Wilson (2011),’An Analytical Approach To Detecting Insurance Fraud Using Logistic Regression’,Journal of Finance and Accountancy.

[4] Lovro Subelj, Stefan Furlan and Marko Bajec(2011),’An expert system for detecting automobile insurance fraud using social network analysis’,Slovene Research Agency ARRS within the researchProgram P2-0359.

[5] Luis Alexandre Rodrigues and Nizam Omar(2014),’Auto Claim Fraud Detection Using Multi Classifier System’,Journal of Computer Science & Information Technology.

[6] Prajna Dora and HariSekharan(2013),’Healthcare Insurance Fraud Detection Leveraging Big Data Analytics’,International Journal of Science and Research (IJSR).

[7] Rosane Vallimand Rodrigo de Mello(2014),’Proposal of a new stability concept to detect changes in unsupervised data stream’,ELSEIVER.

[8] Shin-Ying Huang ,Rua-HuanTsaih and Fang Yu(2014),'Topological pattern discovery and feature extraction for fraudulent financial reporting',ELSEIVER.

[9] Subramaniyaswamy, Vijayakumar, Logesh and Indragandhi (2014),'Unstructured Data Analysis on Big Data using Map Reduce',ELSEIVER.

[10] Yuh-Jye Lee,Yi-RenYeh, and Yu-Chiang Frank Wang, Member, IEEE(2013),'Anomaly Detection via Online Oversampling Principal Component Analysis',IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO.