Middle-East Journal of Scientific Research
24 (Special Issue on Innovations in Information, Embedded and Communication Systems): 185-192, 2016
ISSN 1990-9233; © IDOSI Publications, 2016
DOI: 10.5829/idosi.mejsr.2016.24.IIECS.23159

Restructuring Data and Spatial Relationship Discovery in Detecting Insurance Fraud

N. Pratheeba and N. Dhanalakshmi

CSE Department, PSNA College of Engineering and Technology, Dindigul, India

Abstract: Fraud risk brings a tremendous challenge for the insurance sector. It is a general accord in the market that fraud cases have rapidly expanded in the most recent years. Claims related misrepresentation is the greatest concern towards insurance agencies and the majority of respondents feel that more hostile (anti-fraud) regulations are in great need. Fraud risk in insurance is an unpredictable matter, which affects both the gatherings— insurers and in addition policyholders. Frauds increase the expense of insurance, which results in insurers losing to their competitors and policyholders paying higher premiums. Fraud in insurance happens when people deceive an insurance company or operators to obtain cash for which they are not entitled. Insurance misrepresentations happen each day and in each state. Fraud in insurance also is a crime of violence. Insurance cheaters view insurance fraud as a safe, high-compensate domain. They think fraud is secure and more beneficial than working on other frauds. Insurance fraud is committed by individuals in all different backgrounds. The Insurance Fraud Division (IFD) is trying to fraud in all insurance like health, vehicle, life and other different persons in positions of trust. Additionally, any individual who tries to profit from insurance through making inflated or bogus claims of loss or damage. Almost all insurance agencies effectively fight fraud with Special Investigation Units (SIUs). The investigators regularly have strong law-authorization or insurer claims backgrounds. Insurers also prepare well all representatives to look for false claims. In spite of all, fraud in adversely influencing the insurance division is frequently under-reported or neglected. Subsequently to beat these difficulties, this paper provides machine learning algorithms to consequently identify fraud in insurance companies and the features that cause fraud.

Key words: Big data · Claims · Fraud · Health insurance · Spatial discovery

INTRODUCTION

Data mining is an intense new innovation with extraordinary potential to help organizations concentrate on the most imperative data in the information they have gathered about the conduct of their clients and potential customers. The measure of crude information put away in corporate databases is blasting. From trillions of purpose of-offer exchanges, databases are presently measured in gigabytes and terabytes. The growth of various huge government and private databases has prompted regulations to guarantee that individual records are precise and secure from unapproved viewing or altering.

Also, data mining utilizes advanced numerical calculations to portion the data and assess the likelihood of future occurrences. This can be utilized as a part of direct marketing, trend analysis, interactive marketing and market based examination and fraud detection. This paper focus is on fraud detection to recognize which exchanges are well on the way to be fraudulent and the cases that are non fraud. Organizations lose millions in wages every year since fraud raises their expenses for health coverage and business protection. Fraud in insurance is difficult to identify since many goes undetected. There is sufficient proof to display that insurance fraud is broad and costly. Most insurance agencies effectively battle fraud. Various techniques used to detect fraud is shown in Fig. 1. In any case, a few back up plans still pay certain suspicious cases, trusting it's less expensive than dealing in court. Hence it is important to recognize what is satisfactory to shareholders and policyholders and to strike the right harmony between the expense of preventive action and recognition. One of the difficulties in striking this balance is the way by which one recognizes and measures the genuine expense of fraud cases to an insurance agent.

Corresponding Author: N. Pratheeba, CSE Department, PSNA College of Engineering and Technology, Dindigul, India.

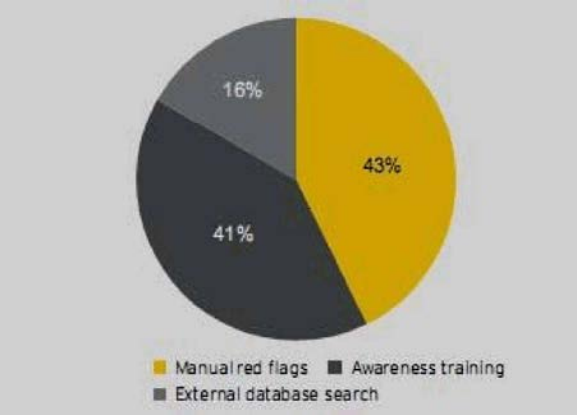

Fig. 1: Means used to detect fraud

Detection of fraud and administration ought to be a proactive procedure, which incorporates recognizable proof of suspicious cases that have a high probability of being fake, through a mechanized measurable analysis. The more the time accessible, the less demanding it is to get ready fake archives. It is hard to do examinations for cases/surrender, which come a while after release from hospital. Rather, data analysis procedures for continuous fraud monitoring can be used, which might distinguish the severity of frauds. The remaining of this paper is organized as follows: Section 2 and 3 briefs about the related works and overview of proposed methodology for detecting insurance fraud Section 4 presents the difficulties of unstructured data and methods to structure them. Section 5 details the fraud and non-fraud classification. Section 6 presents feature extraction for extracting the causes of fraud for data analytics. Finally, Section 7 and 8 concludes this paper with the future work by giving experimental results.

Related Works: In this fragment, we review the works identified with fraud identification. Utilizing SOM and PCA for examining and translating information uses two-dimensional diminishment strategies, Kohonen SOM and PCA for the multivariate examination, investigation and system comprehension of a data set and inside and out examine and translate multidimensional information. It has an unmistakable comprehension and ability to oversee nonlinear issues. The procedure is extremely compelling to stifle multidimensional information sets [1]. A novel cross breed undersampling strategy for mining lopsided datasets in managing an account and protection utilizes krnn (k around closest neighborhood)to perceive and to dispose of the commotion present as oddities in the predominant part class, then one class support vector machine to move bolster vectors in the bigger part class. Yet, it encounters the drawbacks of specificity since affectability is agreed top need in front of specificity [2]. The review on numerous anomaly recognition in information streams is concentrated on here. At first recognizing exceptions in vast information set is a troublesome errand. One strategy that is been done to identify exceptions in vast information set is a technique for distinguishing and erasing separation based anomalies in extensive information sets presents parallel calculation so as with extra time and having eminent execution. At first, a diagram is made in light of the information set. At that point weights are allocated to each of the information's in the diagram. Calculation continues by substituting weights to the information in every column.. By erasing the exceptions, it builds the space for putting away more information [3]. A technique for recognizing accident coverage misrepresentation by social system investigation exhibited frameworks are spoken to in type of systems, for example, social territory, allowing definition and examination of complex relations between elements. Deceitful points of interest are found by using a novel evaluation, Iterative Assessment Calculation (IAA). The system allows the preparation of specialists space learning indeed, even without named information st and it can in like manner be embraced to new sorts of extortion situations when they are observed. The cons of this structure are no accreditation that the limit n/2 is the best choice. It doesn't consider number of sections that has particular marker set and all pointers are managed break even with significance [4]. Multi Classifier strategy for auto Claim Fraud uses a cost network and a blend of classifiers. This work perceives the most cost sparing model to perform the distinguishing proof of related cases with misrepresentation in a dataset of vehicles cases. The structure utilizes Cost Grid blend of classifiers (C4.5, SVM and Naive Bayes computation) to foresee the last request of the each item in the dataset. This work joined the result of each calculation already acquainted with recognize related cases with misrepresentation practices by parallel topology. This is a mild model to perform the revelation of suspected cases with extortion [5]. Using mix of classifiers as a part of a parallel topology makes this framework more compelling. Regardless, capability is not improved since it uses imbalanced classes of information sets. Extortion Detection in human services segment protection utilizing Big Data Analytics, to recognize misrepresentation issues by applying offer information, hadoop environment for quick ID of extortion special cases [6]. Systematic

modules, for instance as choice tree, bunching and Bayesian order are used. The strategy has the ability to distinguish wrong or suspicious records in submitted therapeutic administrations information sets and gives a procedure of how the specialist's office and other social protection data are valuable for perceiving human administrations protection extortion. Unsupervised information streams might have changes regularly. This progressions distinguishing proof relies on upon the surrogate information approach from time course of action examination offers preparing to online unsupervised estimations regardless of the possibility that there ought to be an event of time dependence among perceptions. Yet, it makes no examination of unmistakable division abilities to figure the divergence between PS charts, what might upgrade results for a couple of uses in which repeat assortments are sufficient [7]. Topological example and highlight extraction for fake money related reporting gives a powerful characterization standard to perceive FR in light of the topological examples and an expert forceful component extraction segment to get the striking characteristics of misrepresentation practices. The decided principles are reliant on information and can't be direct associated with other information connections that disregard to satisfy the spatial theory [8]. Map Reduce for unstructured information oversees the tremendous volume of information; manages the huge volume of data. The proposed framework will set up the information in parallel as little lumps and aggregate all the data over the groups in acquiring prepared information. The system has the inconvenience of not executing map diminish work in conveyed mode in which we can use a N number of slaves for a solitary master [9]. Internet oversampling investigation [10] is used to recognize the deviations of special cases from a great deal of information by method for an internet overhauling framework. It chooses the irregularity of the objective specimen. Thereby, cuts the computational costs and memory. However the structure is not best in assessing the principle central headings to handle information in multi dimensional space. By performing this literature study to know the highlights and drawbacks of the existing system [11], we try to overcome the difficulties in the proposed system.

Overview of Proposed Methodology: The flow of the proposed system is shown in Fig. 2. The data set we take will be in unstructured form with all the missing, noise and redundant values. Hence if the data set is taken as such obtained, more time will be wasted on processing them and more memory space will also be wasted.

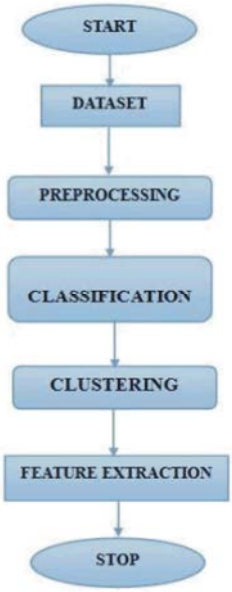

Fig. 2: Work flow of the proposed method

By organizing the data in correct format, it can distinguish fraudulent records in submitted health care data sets and gives a methodology of how the clinic and other social insurance information is useful for the identifying medicinal service insurance fraud. Hence preprocessing steps are carried out and unstructured data set is converted to structured form for simplified processing. Yet, the data set may contain some outliers (i.e.) the data that are not relevant to our data set. Outliers are also known as anomalies. The outliers are detected based on neighborhood distance and are removed. Classification is performed to classify fraud and non-fraud data. In correctly predicting both fraudulent and genuine claims, both sensitivity and specificity should be given equal priority .Fraud data set is alone extracted and features that cause fraud are identified. Thus the proposed approach will determine the spatial relationship of fraudulent records and extract the characteristics that cause fraud. Data analysis is done on the clustered fraud data set and to know the characteristics and features that cause fraud. Finally performance is measured for the proposed system. Hence this paper focus is on to improve the performance of classification and clustering by considering both sensitivity and specificity.

Structuring Data: Unstructured Data alludes to data that either does not have a predefined information show or is not composed in a predefined way. Unstructured data contains information, for example, dates, numbers and

A	B	C	D	E	F	G	H	I
2	PID002	Balaji	M	32	1	ARUN	HOSPITAL	5000
3	PID003	Jerry	M	40	0	ARUN	HOSPITAL	2500
4	PID004	Robert	0	35	1	KOVAI	HOSPITAL	7500
5	PID005	Priya	F	22	1	RAJA	HOSPITAL	4500
6	PID006	David	M	25	1	RAJA	HOSPITAL	4500
7	PID007	Balaji	M	32	1	ARUN	HOSPITAL	5000
8	PID008	Jerry	M	40	0	ARUN	HOSPITAL	2500
9	PID009	Robert	0	35	1	KOVAI	HOSPITAL	7500
10	PID0010	Priya	F	22	1	RAJA	HOSPITAL	4500
11	PID0011	David	M	25	1	RAJA	HOSPITAL	4500
12	PID0012	Balaji	M	32	1	ARUN	HOSPITAL	5000
13	PID0013	Jerry	M	40	0	ARUN	HOSPITAL	2500
14	PID0014	Robert	M	35	1	KOVAI	HOSPITAL	7500
15	PID0015	0	F	22	1	RAJA	HOSPITAL	4500

Fig. 3: After Data Structuring (Zero Substitution)

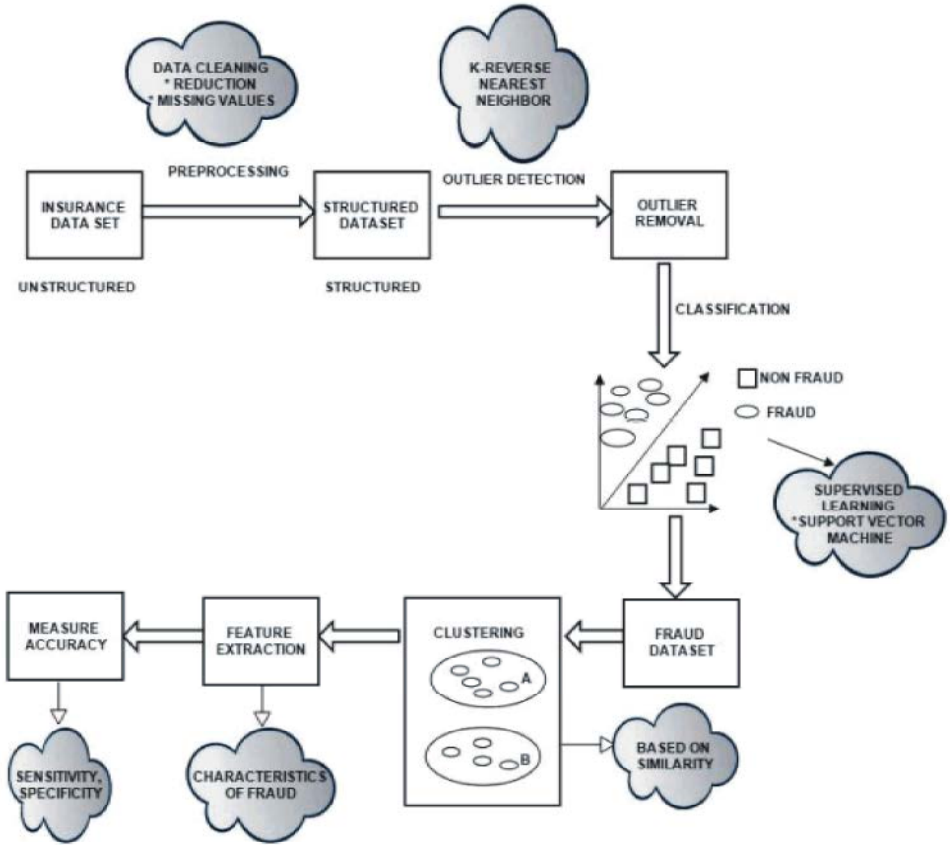

Fig. 4: System Architecture

actualities also. This outcome in anomalies and ambiguities that make it hard to interpret the data. It has all the missing qualities, subsequently missing the estimation of that information is not that much simple since information can't be analyzed, searched, sorted and imagined. Substantial measure of these data needs systematic organized plan for handling the information. Data cleaning is gone through procedures, for example, filling in missing rows/columns, noisy data smoothing or determining the irregularities present. Filling missing data

in rows/columns should be possible by zero substitution as shown in Fig. 3 missing qualities can be simply loaded with zero to make segments equal. Else mean value substitution technique in which mean estimations of all segments is substituted in the missing spot.

Also sometimes, attributes may have repeated values. So it may be reduced to speed up the computation and to free the memory space. For example, the attribute age with numerical values is removed. Furthermore, the attributes Year, Month, Week and Day represent the date of the accident and the attributes Month claimed, Week claimed and Day claimed represent the date of the insurance claim. Thus, a new attribute Time Gap can be derived from five attributes such as Year, Month, Week, Day and Month claimed. Subsequent to preprocessing is over data set is checked for exceptions (outliers) as shown in system architecture, Fig. 4. Outlier is a perception point that is removed by different analysis methods. An anomaly might be because of variability in the estimation or it might demonstrate trial blunder and hence it may be rejected from the data set. Anomalies can happen by chance in any appropriation, yet they regularly show estimation mistake. This can be because of coincidental precise mistake or defects in the hypothesis that created an expected group of likelihood appropriations or it might be that a few perceptions are a long way from the focal point of the data.

Outliers focus in this manner show defective information, mistaken systems or ranges where a specific hypothesis won't not be valid. In huge data, a little number of exceptions that are not ordinary may occur in common. There might have been a mistake in data transmission or interpretation. Anomalies emerge because of changes in framework error, false conduct, human blunder, instrument mistake or essentially through normal deviations. For this purpose and to reduce the time required for classifying the not needed data, outlier detection and removal is used. The algorithm used for this purpose is given below.

K-Means for Anomaly Detection: K means using Nearest Neighbor calculation is utilized for anomaly identification. Let X be d-dimensional information set, p1, p2; p3;… ; pi; pj;… ; pn, where n is size of the data set and pi; pj are any two focuses (tests) in X. sdij is the separating length between two points pi and pj.

Input: data set S = (U, A, V, f), where U = n and A = m; threshold tl.
Output: a set O of neighborhood-based outliers.
For every a 2 A {
For every x 2 U {
Calculate dist(xa, xu);
}
}
For every x 2 U {
For every y 2 U {
For every a 2 A {
Calculate dist(xa, ya)
}
Calculate Neigh(x), Neigh(y);
}
if Neigh(x)
If Neigh(x) > l, then O = O [x];
}
Return O.

The arrangement of krnn of point pi gives an arrangement of focuses that consider pi as their k-closest, for a given estimation of k. On the off chance that a point xi has higher number of krnn than another point pj, then we can say that pi has a denser neighborhood than pj. As it were, less the quantities of krnn, the more distant separated are the focuses in the dataset to pj, i.e. the area is sparse.

Fraud/Non-Fraud Classification: After all the anomalies are uprooted, the dataset is under supervised learning. The organized dataset is gone as a data to classification as in Fig. 5.

The algorithms work by building a model from illustration inputs with a specific end goal to settle on data driven forecasts or choices. Supervised learning is the machine learning undertaking of surmising a capacity from labeled trained data.

The trained data comprise of an arrangement of training illustrations. Every illustration in supervised learning is a pair comprising of an given value and a yield esteem. An ideal situation will take into account the calculation to accurately decide the class names for new values. This is used to generalize from training dataset to new data by supervised learning. Hence the system is given with illustration inputs and their wanted yields and the objective is to obtain a general deciding rule that maps inputs to yields. Example of trained dataset is given below.

David, M, 25, 1, RAJA HOSPITAL, 4500, legal
Balaji, M, 32, 1, ARUN HOSPITAL, 5000, fraud
Jessy, F, 40, 0, ARUN HOSPITAL, 2500, legal Robert, M, 35, 1, KOVAI HOSPITAL, 7500, legal Priya, F, 22, 1, RAJA HOSPITAL, 4500, fraud

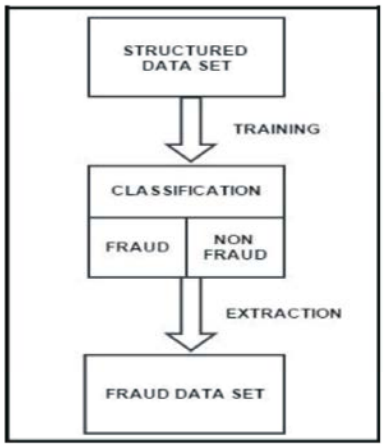

Fig. 5: Fraud Dataset Extraction

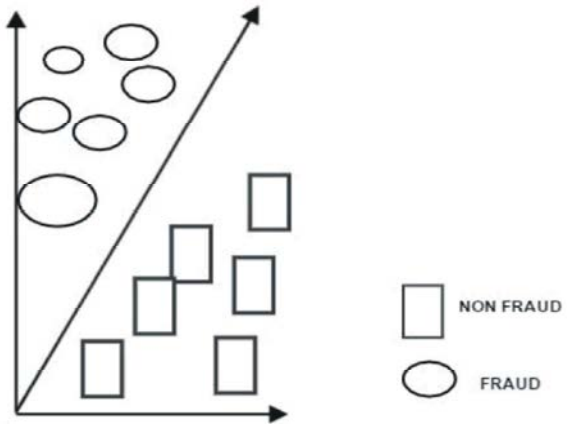

Fig. 6: Fraud/non-fraud classification

Prakash	M	22	1	RAJA HOSPITAL	4500

Fig. 7: Testing Data to Classify

Here in this example, three cases are legal and 2 cases are fraud. Hence by calculating the probabilities we get legal: 3/5=0.6 and fraud: 2/5=0.4. An example in the patient database rule is If Age is between 20 and 25 and disease type= diabetes and no. of methods covered is >5 and length of stay is >5 then it found that as an abnormal case. And same as in insurance claim dataset if insurer is from region 'aaa ' having disease liver-surgery and have diagnosis procedure coding error =yes and delay in submitting the document =yes and so on then it's a possibility that his claim going to be rejected.

Support Vector Machines (SVM): After training the data set, support vector machine is utilized for classification into fraud and non fraud data set. It is a mining technique (machine learning) system used to foresee the prediction. For instance, you might wish to utilize classification to decide whether the fraud on a insurance data will be fraud or non-fraud. Grouping comprises of anticipating a specific result taking into account a given data. With a specific end goal to foresee the result, the algorithm forms a training set containing attributes and their respective results. For instance, in a therapeutic database the training set would have significant patient data recorded beforehand, where the forecast quality is regardless of whether the patient showed some heart issue.

IF (Age=50 AND Heart rate>80) OR (Age>65 AND Blood pressure>140/70) THEN Heart problem=yes

SVM are supervised learning models with related learning calculations utilized for characterization and repeated investigation. Given a dataset of trained cases, each set apart to belong to one of two classes, fraud and non-fraud data. This algorithm constructs a model that maps new cases into one classification or the other, making it a non-probabilistic binary classifier. New examples are then mapped into that same space and anticipated to have a place with a class in light of which side they fall on. All the more formally, SVM develops a hyperplane in a high or unbounded dimensional space as in Fig. 6 which can be utilized for characterization, classification, regression. Hence, bigger the edge- lower the speculation blunder of the classifier. Consider the testing data in the table to perform classification as in Fig. 7. It is to test whether the data in the record is fraud or non-fraud.

For the testing data, find probabilities by calculating the number of occurrences divided by the total number of occurrences. First we find probability for non-fraud data. Therefore we obtain the probability for non fraud data as 0.0032. Similarly, we obtain for fraud probability as 0.0016.now check for which the probability is greater. From the obtained results, we know probability for non-fraud is greater than probability for fraud. To keep the computational burden sensible, the mappings utilized by SVM are intended to guarantee that dot products might be processed effortlessly as far as the variables in the first space, by characterizing in terms of a kernel function k(x, y) selected to suit the problem.

Spatial Hypothesis: From the classified dataset, fraud set is separated from everyone else extracted as data and clustering is executed. Clustering is then performed on the component space. Clustering will frame bunches with comparable qualities of fraud. It is a division of information into gatherings (groups) of similar features.

When data is represented by only fewer clusters, it may necessarily lose certain fine details though it aims simplification. It shows information by its groups.

Feature Extraction: By performing feature extraction, attributes that cause the insurance fraud can be identified. There may be numerous features that cause fraud. Grouping them based on their similarities will help us to identify the major causes of the fraud so that we can be cautious of those frauds by preventing them in future. This extraction begins from a initial set of quantified data and derives determined qualities (features) expected to be knowledgeable, non repetitive, encouraging the learning and speculation steps, now and again prompting better human elucidations. Based on the algorithm above, when applying feature set Fj using Cluster Feature Set (CFS) on all attributes of the given record features that cause fraud are detected. Finally performance analysis is done to find the efficiency.

Experimental Setup and Results: Beginning with a data set of over 500 providers, the set was narrowed to roughly 360 providers through selection criteria. After performing the analysis, only 35 providers raised 2 or more of the potential 14 predictive flags. 17 providers raised 3 or more flags. We focused on these 17 providers to evaluate the potential efficacy of the approach. We interviewed qualified health care fraud subject matter experts to evaluate the claims of and the raised flags by these 17 providers as given in Fig. 8 and Fig. 9. While some of the flags could be understood as acceptable given the types of services rendered or due to the provider's operating environment, there was a preponderance of evidence suggesting that at least 12 of these 17 providers (71%) with three or more flags should be immediately referred for audit and potentially to law enforcement.

When outliers are detected by forming clusters, those anomalies are detected and removed by nearest neighborhood algorithm. It utilizes k-means method for this purpose. Hence by removing the outliers the large deviations from the dataset are removed for easy calculation and to reduce time and increase efficiency. Later by svm classification is performed, if the kernel returns positive value it is recognized as normal data else fraud data as in Fig. 10.

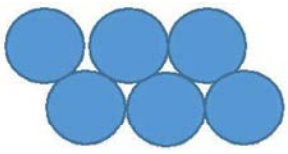

Fig. 8: Claim Nodes

Fig. 9: Feature Extraction

```
21 3.3166247903554
22 5.0
23 5.0
24 5.0
25 5.0
26 18000.02524998229
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Outlier is : 26
Outlier is : 18000.02524998229
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
-0.8144038847128741 : Fraud
0.8428025910218366 : Normal
1.3720942506154188 : Normal
2.4773672602828163 : Normal
-3.6751309527496687 : Fraud
-0.8144038847128741 : Fraud
```

Fig. 10: Claim Nodes

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 fraud detection techniques to classify the records into fraud and non-fraud. The analysis methods which will be helpful in the field of health insurance is performed. The strength of this work is not only in identifying the fraud data, but also it helps to identify the features that are the reasons for fraud to take place. Hence it can be utilized for future occurrences to easily detect fraud without much more time. The future work can be extended to all types of insurance fraud which will be helpful to the society in lot of ways.

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 undersampling method for mining unbalanced datasets in banking and insurance', ELSEIVER.
3. Beigi, M.S., S.F. Chang, S. Ebadollahi, D.C. Verma, 'Anomaly detection in information streams without prior domain knowledge', IEEE Xplore, Ibm Journal of Research and Development.

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 research Program, pp: 2-0359.
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 Hari Sekharan, 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-Huan Tsaih 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, 2013. 'Anomaly Detection via Online Oversampling Principal Component Analysis', IEEE Transactions on Knowledge and Data Engineering, 25(7).
11. Pratheeba and Dhanalakshmi, 2015. 'A Typical Study of Improving Accuracy in Detecting Insurance Fraud on Unstructured Data Sets', International Journal of Engineering Sciences and Research Technology (IJESRT).