Jestr

Journal of Engineering Science and Technology Review 13 (1) (2020) 204 - 213

JOURNAL OF
Engineering Science and Technology Review

Research Article

www.jestr.org

A Supervised Approach for Detection of Outliers in Healthcare Claims Data

P Naga Jyothi[1,*], D Rajya Lakshmi[2] and K.V.S.N.Rama Rao[1]

[1]*Dept.of CSE, K L Educational Foundation,Guntur,A.P, India*
[2]*Dept. of Computer Science and Engineering, JNTUK UCEV,A.P, India*

Received 18 March 2019; Accepted 10 December 2019

Abstract

Outlier detection is a fast-moving method in healthcare data and it is the major concern for the health insurance providers. Most of the Medicare data is related to real-world data. Outlier analysis plays a crucial role in data validity and reliability. To detect outlier for medical data is a complex task as it is having more number of variables and is of multivariate in nature. The paper presents a model-based approach in which outliers are detected and they were assigned with labels. The outlier or suspicious is defined as some outcome, which is expected that it is going to commit fraud. The methodology carried out in two phases to develop a Supervised Outlier Detection Approach in healthcare Claims (SODAC). Initially, the data preprocessing stage for feature selection it uses the filter method and set grouping hierarchy to select the best subset and to organize the features. The outlier detection phase uses the combination of classic methods of statistical and distance-based approach. To evaluate the distribution of data the Gaussian probability density function is applied for the data values. The distance-based approach which reflects the outputs as outlier codes. The partitioning of the input dataset and applies statistical mean to each subset and further uses derived multi aggregate metric to consolidate the data instances of the partitions(subsets). The distance-based outlier detection (dod) is done by calculating the maximum distance from the inner average statistical mean measure of the neighborhood from the data objects (instances) of the input. The data object value goes beyond the maximum or minimum of calculated measure is termed as suspicious. This type of detection of outliers is called as indicative fraud potential. The results performed relatively stable for a large scale data as illustrated in the experimentation part over publicly available real world data.

Keywords: Detection, Healthcare, Data, Supervised Method

1. Introduction

In the fields of data mining and statistics, outlier detection is one of the tasks. The outlier is an observation where its value deviates from a normal and notifies out as suspicious. In the process of data analysis, to identify and reveal the systematic errors in the data, in which the outlier analysis primarily works. Outlier analysis is gaining importance in wide applications domains like a tax, credit card, insurance, cyber security, military, healthcare, etc. The strategies of outlier detection (Fig. 1) is to first define the normal region (behavior) to every possibility considering all factors and, it is tricky. Defining the boundary of normal to outlying data is very ambiguous because there is a slight change between these two points is defined by Varun Chandola et.al [1]. By considering various constraints and requirements the designs of outlier detection models are different for different applications and which are very specific to the domain. Tan P et.al [2] gave the factors need to consider for outlier detection is the nature of data, application domain, and its knowledge discipline. The outlier detection technique is to find the abnormal patterns from the input data. The nature of input data instance has many attributes are of different types like categorical, binary and may be multivariate or univariate (multiple or single data types). The important feature of outlier detection technique is to select the best features from the input data for the algorithm to give the best results. Apart from the nature of data, most predictive models have been using the labeled data for training purpose in which labels generally define the normal or outlier.

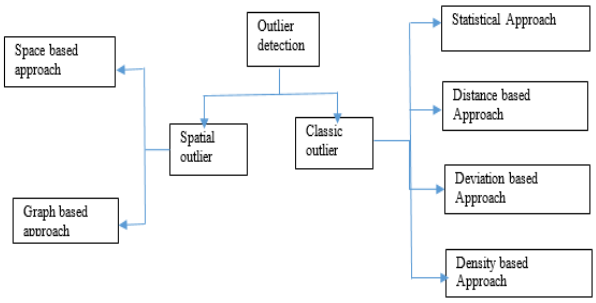

Fig. 1. Outlier detection approach

1.1 Classification based on the type of supervision

The techniques are categorized into three Supervised, Unsupervised and Semi-Supervised outlier detection techniques which are relied on the extent the labels are used by Mitchell T M [3] and Vapnik V N [4].

i. Supervised outlier detection Technique: As per Abe et.al [5] it is an approach to build an accurate predictive model as data instances are fully labeled and can be categorized as an outlier or normal. To get labeled data is often expensive as it is collected from the human experts so it requires more effort.

*E-mail address: pbtjyothiraj.33@gmail.com
ISSN: 1791-2377
doi:10.25103/jestr.131.25

ii. Unsupervised outlier detection Technique: The approach in which, to make assumptions about the data like frequently occurring data instances as normal otherwise outlier. This technique suffers from high false rate, as assumptions do not hold true discussed by Varun Chandola et.al [1].

iii. Semi-supervised outlier detection Technique: Semi-supervised techniques assume labeled data instances in one class and it difficult to label data from other class. It is an approach to model normal instances during training and it is difficult to find every possibility of outlying behavior that is present in data is limited is briefed by Markov et.al [7] [8] in two approaches.

Based on real-world applications the use of the technique, method, system or model varies. Likewise, the comparison of those outlier detection algorithms must rely on some specific measures which test the quality of the technique. Some of the studies of Kamaljeet Kaur and Atul Garg[9] discusses the brief view over empirical analysis of outlier detection algorithms. As part of their conclusion Clustering and Classification, based outlier detection algorithms are efficient and highly scalable, the computational cost is low. Whereas, in contrast, the statistical and depth based, sliding window are at the lowest level. The distance-based and DSS outlier are efficient, low computational cost and can be extended to large dimensional data.

1.2 Importance of metrics

To illustrate any case study for identification of fraud the fraud metric extraction must be defined at the prior stage. Metrics can be designed or derived for the particular application domain. Metric identification requires knowledge from experts of domain and statistics and is a complex task. Filtering of non-fraud to fraud indicators will normally have a statistical deviation from a group of normal data. Our model devised a derived multiple aggregate metrics to summarize and analyze the complete dataset. The metric can be defined as applying an aggregate function to two or more columns at the same time or two different (multiple) aggregating functions can be applied on the same column.

1.3 Outlier's detection in healthcare data

The definition given by Laurikkala J et.al [10] for outlier, which is to be detected will be an input to the outlier detection technique. Each data instance is analyzed with respect to remaining data instances and defines it as outlier i.e., based on the inconsistent values of an attribute. This type of outliers is defined as Type-I outliers. Type-I are defined for applications in the works of Ghosh et.al[11] like credit card fraud detection, medical records regarding the patient health, Insurance claim fraud detection, mobile phone fraud detection etc., The outliers are detected based on the insurance claiming amount of provider and identifies as suspicion. The outcome is expected to detect fraud instances and not the providers who are committing fraud given by He Z et.al [12].

The works of R J Bolton et.al[13] and E Elmer et.al[14] discusses the behavioral perspectives of health insurance data are desirable for designing metrics in the detection process of outlier or fraud instance. Health insurance environment is of a complex nature and consists of large data. The detection of outliers is challenging because it involves various organizations and departments each of which targets at a different geographic level. Different schemes exist for different perspectives, and these allow us to determine the utility of finding fraud at each level. To deal with various characteristics of data like volume, complexity, diversity, and reliability data labels the model must able to choose the right metrics and associated mining technique. Derived multiple aggregate is the proposed metric which is used to categorize the feature claim amount for a particular procedure undergone at a specific location.

Therefore, the paper incorporates preprocessing and Histogram-Based techniques to visualize the instances of outliers which is drawn from works of Guido et.al. [15].To investigate the outlier to be specific it needs to incorporate some of the domain-specific metrics. The outlier's techniques use, classical approach of statistical model and distance based mainly briefed from the works of Markus G et.al [16], Xi Jingke [17] and Indukuri BhaskaraRaju [18].

2. Knowledge base from Related Works

The work of many researchers is related to the healthcare systems majority of the works firstly, defines metrics for data analysis and application of cluster-based or distance based algorithms on the particular disease and finding abnormalities or outliers. Secondly, finding fraudulent activities i.e. overbilling for services with an association of providers and physicians or patients. In this perspective, this work presents a methodology for finding outliers based on the claim amount which was provided by the hospital management for performing the surgery.

The detection of fraud based on the outlier in health care systems of U S Medicaid data on which most of the researchers had performed their results. According to Dallas Thornton et.al [19], proposed specific metrics for finding fraud which is claimed by the dental providers. Regression analysis is done between the amount reimbursed by Medicaid to the provider and the number of reimbursed claims. His work developed anti-fraud communicated a model to stakeholders to make a threat to fraud. From the work of Milou Meltzer [20], detection of outliers with mixed attributes by using Attribute bagging added value with K-medians clustering in Wisconsin Breast Cancer dataset. The infrequent pattern analysis is performed on the continuous data based on support or frequency and calculates the outlier score. To avoid the curse of dimensionality problem the high dimensional categorical data is transformed into multiple binary values. The work of Jian Wu et.al [21] presents detection of medical fraud in medical insurance audits with improved outlier detection algorithm based on K-Means clustering on pulmonary heart disease, pneumonia, and chronic bronchitis disease datasets.The paper of Chirsty.A et.al[22] proposed two algorithms for removing outliers using the outlier score of high dimensional data in order. The approaches of Cluster-based outlier detection algorithm and Distance-based outlier detection were used to construct the framework for efficient patient care and disease assessment constraints for Esoph, Diabetes, and Kostecki Dillon datasets. Medical fraud inclines to be more crucial for medical insurance providers in China from the paper of Weijia Zhang and Xiaofeng [23] proposes improved local outlier factor (imLOF) a spatial density algorithm for anomaly detection of illegal health claims. The extension of work is intended for more labeled data and development supervised model. From the work of Varun Chandola et.al [24] the knowledge discovery is carried with the transformation of transactional data, identifies the fraudulent providers from the real claims data. The properties of the provider network, are the extraction of a relevant feature from the provider network, relevance for identifying

the fraud. The analysis of claims sequences is done by temporal analytics by applying the change of point detection with statistical process control and anomaly detection by CUSUM statistic which is based on the number of claims submitted by the provider. Archana Kadam and Sagar G Powar [25] proposed a hybrid approach for outlier detection in the medical dataset. A density-based with k-means partition technique is used to estimate the outliers of heart disease data. The survey of Yijun Lu [26] and Xuanfan Wu [27] gives a brief view of the metrics, techniques, and tools of anomaly detection strategies and their applications, advantages, and disadvantages. According to Peter Travaille et.al [28] electronic fraud is also relatively increasing in large industries of banking, telecommunications, insurance, automobile sectors etc. The investment towards fraud controls comparatively low as the claim systems are very complex w.r.t to healthcare. The use of statistical methods by Tang et .al [29] of data mining approaches shows outstanding performance in order to explore the healthcare field for outlier detection. The outlier score determines the fraud in the prescription provided by provider group. The prescription based outlier detection is discussed by Tsoi et.al [30] by the following data mining techniques with feature selection, clustering, pattern recognition and outlier detection.

3. Sources of Referential Medicaid Data

The privatization of health insurance is introduced in India in 1999. The government of Andhra Pradesh is introduced in 2007, Rajiv Aarogyasri Scheme (RAS) health care trust for the welfare of the public. The annual authorized expenditure towards health care under this scheme is of 925 hundred crores. It is the largest scheme in the world and covers 50 million people and with more than 360 providers. The maximum percentage i.e. 71% of networked hospitals is private. After bifurcation of Andhra Pradesh, it transformed to NTR Vaidya Seva Scheme (NTRVSS). In 2013 which covers 13 districts with 1044 procedures and 138 packages. The financial coverage of each family from 2 to 2.50 lakhs. This scheme covers 1, 59, 95,433 families which are under Below Poverty Line (BPL) of state. The variations of package rates for the procedures in which they undergo are based on intercity, inter schemes and city wise were dictated by the state policy were discussed by Indukuri BhaskaraRaju [18]. The source of dataset is publicly available [31].

Medical providers have a unique role in claiming the amount and therefore, they are the beneficiary of such outcomes. The health insurance is marked as a large number of fraudulent claims. To detect such irregularities, there must be an approximate estimation for the procedure so that, it can be overcome. Therefore, there must be geometry coverage of hospital location, the severity of disease and level of treatment and coverage of scheme all these should be considered given by Markus G et.al [16].

4. Methodology for detection of outlier in healthcare claims

The recommended methodology enables the successful identification of fraudulent claiming behavior. The outlier predictors use labels to explore the suspicious claim amount. We have developed a two-phase model iterative process to identify the outlier in the healthcare claims data. The subsequent subsections describe each phase of this methodology, and for experimentation, it is carried under the environment of Python.

4.1 Feature selection

Feature selection, for heterogeneous columns and categorical kind of datasets plays an important role. This process enables the algorithm to converge and train faster, reduces complexity, easier to interpret and improves efficiency if the right subset is chosen finally, enables over-fitting. For experimentation purpose, the paperwork has applied filter methods (Fig. 2.) to select the subset of attributes. The filter generally is independent of any machine learning algorithm at the preprocessing stage. Instead of using scores to select the features based on statistical tests for their correlation is subjective with the outcome variable. The selected set of features are prompt for the outlier detection of the claimed amount for surgeries for various categories undergone in a different hospital at various locations. The relativeness and dependencies of each feature are considered and the transformation of some features was done for the purpose of experimentation.

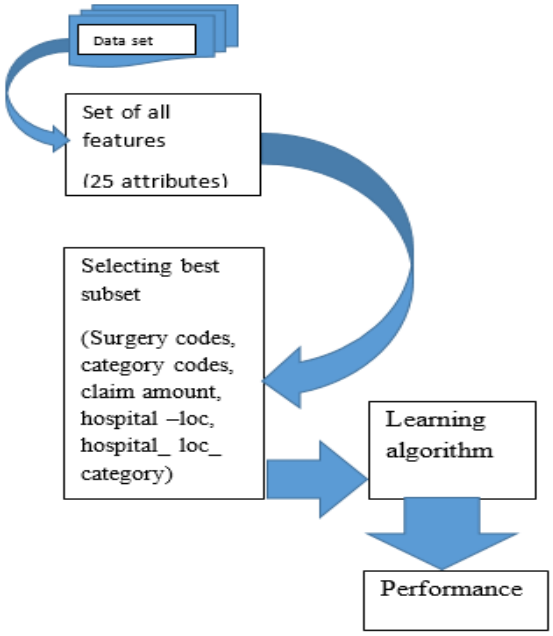

Fig. 2. Filter method feature subset selection

The attribute category-code which defines the set of surgery codes, surgery–code defines the exact surgery which is done to the particular patient. The claim amount is defined for various surgeries codes at a different location in various ways at a different hospital. The nature of NTR Vaidya Seva Scheme data is observed that each record is significantly unique.

4.2 Application of Preprocessing Technique for NTRVSS dataset

The collected healthcare data is of a robust set of heterogeneous type and all of the categorical attributes except few like discharge-date, surgery-date, claim-date, and card number and claim amount (Table. 1). So, while processing data for analysis it addresses a variety of data quality problems as they are weakly related to each other. This paper mainly focus on the claim amount of surgery which had taken place at the hospital at a particular location. The data of each cardholder is probably claiming different claim amount for different surgeries that they have undergone. So, most

probably the data do not possess any kind of similarities in the claim amount for the same surgery at the same hospital. This work proposes an estimation of the maximum amount that can be claimed for the particular surgery at a different location. The supervised outlier detection technique based on the previous instances or records detects overbilling of claims and results as a fraudulent claim. In order to proceed for experimentation as all the records are unique. The techniques of preprocessing are either filling the missing values, reducing the dimensions, normalizing the data dimension or removing the redundancies do not suit this kind of healthcare dataw.r.t to surgery code and hospital location code. The derived multiple aggregate metric is made use from the pandas package of python for experimentation purpose.

Table 1. Summary of features used in NTRVSS dataset

Feature number	Description
1-2	Patient Identification
3--5	Patient personal Information
6-9	Diagnosis procedure codes
10-12	Demographics of patient
13-14	Previous cost history
15-16	Claim history
17-20	Diagnosis demographics
21-22	Date of surgery and discharge
23-24	Mortality details
25	Source of registration

This paper adds a feature or attribute to the dataset so as to bring the data to be more general. The transformed dataset will definitely give better results when compared to the whole dataset records as individual data instances. The generalization using concept hierarchy of the dataset is done w.r.t Hospital location category as RURAL, SEMI-URBAN, and URBAN as three different class labels. The mapping of the whole dataset is done by adding a new column or attribute and reducing each record to as RURAL, SEMI-URBAN, and URBAN with reference to Wikipedia. The background knowledge of the data set is defined with one of the key concepts of schema hierarchy.

4.2.1 Concept hierarchy

A concept hierarchy H is a poset (partially ordered set), (h, <), where h is a finite set of concepts, and < is a partial order on h.

4.2.1.1 Set grouping hierarchy

This is type of concept hierarchy is defined as grouping relationships into a set of concepts (values of attributes) in order to reflect the semantic relationships characteristics in the given application domain. It is called as instance hierarchy as it works on partial order of hierarchy on a set of instances or values of attributes for the structured attribute as it is more of operational sense. This can be used for altering a schema hierarchy or one more set grouping hierarchy to form a refined hierarchy from the study of Christy et.al [22]. The transformed dataset relatively works for finding the suspicious claiming behavior of healthcare data (Fig. 3). The usage of metric is to analyze NTRVSS data set, i.e., by calculating aggregate summarized claim amount.

Define hierarchy Procedure-codes as
level3 (Hospital-Location-category) :{ URBAN, SEMI-URBAN, RURAL} < Level 2 : surgery-codes
level2 (surgery-code):{M1.1,M1.2…….M12.9,S1.1.1,S1.1.2……..S17.1.4.1}<level1:category-codes;
level1 (category-code): { M1 ,M2 ,……M13,S1,…..S17} < level 0:all-category-codes.
Definition of set grouping hierarchy

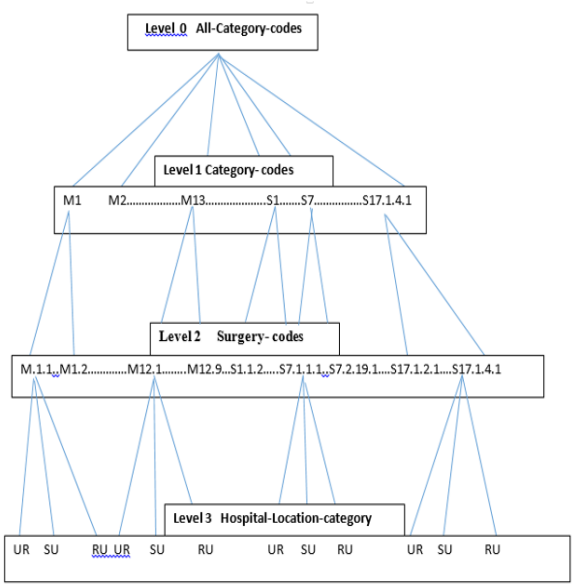

Fig. 3. Schema hierarchy using set grouping hierarchy definition for Schema hierarchy

4.3 Proposed Algorithm for Supervised Outlier Detection Approach in healthcare Claims (SODAC)

The recommended system allows the identification of fraudulent behavior in medical claim and associates to patient or professional submitted to the organization. The analyses of different features and their cross-references based on the number of times of claiming amount at different locations for various surgeries.

Initially, each claim is determined by its category code and surgery code and evaluates the claim amount w.r.t different hospital location codes. The normal distribution of each claim amount is calculated by Gaussian probability density function which has been performed at the previous stage and is considered for further outlier detection.

The probability density function of Gaussian distribution is given

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1)$$

The outlier code is evaluated by distance-based outlier detection (dod) is formulated as

Definition: Let N_p, is set of all instances of D, d_1 is subset of all instance values w.r.t to an object X_p. The value of X_p is determined by the average statistical mean measure from all set of objects in instances values w.r.t to A_p. $X_p \in (x_1, x_2 \ldots x_i \ldots\ldots x_p)$, let $avgdist(x, \bar{x}) >= 0$ be the measure of distance between the instances x and $\bar{x}$

The Statistical mean measure for object X_p is defined as

$$\overline{X_p} = \sum_{m=1}^{p} \frac{x_m}{n} \quad (2)$$

If X_p is calculated w.r.t to an object $N_{p,}$ *k and* k_1 are the number of instances values of D and d_1.
The average distance of $N_{p,}$ of X_p is calculated by the formula

$$\bar{D}_{x_p} = \frac{1}{k} \sum_{x_i, x_r \in N_p} \frac{X_p}{k_1} \quad (3)$$

The inner distance of all instances of X_p for an object x_p is evaluated by the normal deviation, maximum and minimum values of $X_{p.}$ The distance-based outlier is determined by the

$$dod_{x_p} > \max(\bar{D}_{x_p})$$

$$dod_{x_p}, < \min(\bar{D}_{x_p})$$

For instance dataset D, the k are instances of category codeM1, M2,……..., M12 and S1, S2,……..., S17.Let of M1 be the category code where it retrieves nearest claim values of different surgery code ranges from {M1.1, M1.2……..M1.7, M.1.8} as sub-instances k_1.The claim of surgery code M1.1 compares with all instances of the whole dataset and take the local claim amount. The claim amount M1.1 w.r.t to each hospital-loc-category {Rural, Semi-Urban, Urban} is also evaluated. The outlier claim amount of surgery is calculated by comparing the claim amount w.r.t to the previously stored values and identifies the fraud claiming behavior. The instances of $dod_{M1.1} < dod_{M1.1r}$, which sorts the objects according to their dod values. The highest values of dod for each surgery code is termed as maximum claim amount. The amount which exceeds this maximum and undergoes to minimum amount they are termed as outlier cost.

5. Experimental Results

In this section, the outlier detection is carried with a real-world dataset which is provided by the government of Andhra Pradesh. The paper conducted experiments using a Python environment using pandas, numpy, matplot and scikit packages. The experiment uses two baseline algorithms for detecting the outliers. The Statistical mean measure recalls the first phase of experimentation. The distance measure which appears at outlier detection stage improves the performance of the algorithm in identifying the suspicious behavior of patient or professional. The subsections discuss the dataset description and result at various phases of experimentation.

5.1 Medical claim Dataset description (NTRVSS)

The preliminary level of any algorithm is data preparation and feature selection which have an important aspect for data analysis and these directly rely on the quality of output. For the data analysis task, the domain-specific features which should be confined to domain knowledge definitely take much time to select and define. The selected features for analysis is necessary to apply data transformation, data reduction, normalization or is it require defining explicitly new features for processing. This work has collected data from NTR Vaidya Seva Scheme. The dataset is defined with 1,60,949 records and 25 attributes as Card number,Card-Midnumber,age,sex,caste,category-code,category-name,surgery-code,surgery-name,village,mandal,district,pre-authorized,pre-authorize-amount, claim-date, claim-amount Hospital-name, Hospital–type, hospital location, hospital district, surgery–date, discharge-date, mortality, mortality-date and source-of-registration. The attributes are of a heterogeneous type, categorical data and having multiple labels for each data instance.

5.2 Results and Discussion

Firstly, experimental dataset undergone preprocess stage by using feature subset selection and set grouping hierarchy concept (see from Figure 3.) for schema restructuring and which results in the transforming the dataset (see in Table 1) to resultant subset features shown in Fig 4.

1. Initialize claim amount of sur_code, hos-loc-cat, outliers codes
2. for sur_code in Surgery_code
   begin
   val<-length of Mean_data[sur_code]
   for h_loc in range(val):
   begin
   if Hospital_loc[h_loc] in Mean_data[sur_code]:
   then
   if max_claim_amount[i]> int(Mean_data[sur_code][Hospital_loc[h_loc]])
   then
   if sur_code in Outliers
   then
   assign new_ Outliers <- {Hospital_loc[h_loc]: max_claim_amount[i]}
   Outliers [sur_code].append(new_Outfilers)
   else:
   new_Outliers<- {Hospital_loc[h_loc]:max_claim_amount[i]}
   Outliers[sur_code]<- [new_Outfilers]
   sur_code_temp_list.append(sur_code)

   temp_loc_list.append(Hospital_loc[h_loc])

   temp_max_amount.append(max_claim_amount[i])

   mean_data_list.append(Mean_data[sur_code][Hospital_loc[h_loc]])
   i<-i+1
   else:
   val<-val+1
3. Sort all the outliers according dod based on surgery codes w.r.t hos-loc-cat in a dictionary format.
4. The highest values and lowest of dod for each surgery code is termed as outliers codes.

Pseudo code for outlier detection using SODAC approach

Next to applying of Gaussian pdf values (shown at formula (1)) of various surgery codes at a different location in order to calculate the distribution of data. The following observations shown in Figure 5.are of minimum, normal and outlier distribution of claim amount for each at particular surgery code w.r.t different hospital locations. This section moves on to further experimentation for outlier detection using distance and statistical measure. The techniques of a supervised model are that the data have undergone training to the whole dataset in the previous stages. The average statistical mean of each sub category of surgery code is

calculated by the given formula(2) and (3).The performance of the metric for our method recalls the values of each surgery code w.r.t to their claim amounts along with newly added column i.e. hospital-location –code. The Figure 6 shows clear picture of how a derived multi aggregate metric works for the calculation of complete dataset. Similarly, the individual list of table were shown in Figure 7 as we see the huge change of claim amount. The same replicated in console output in Figure 8 shows, all the outliers codes for all surgeries. For understanding the comparison of claiming amount, at different hospital location. The working principle of an algorithm for detection of outlier codes, we have collected some observations of the experiment in an excel format it is shown in figure 9, for comparison of initial to the output. The Figure 10 and Figure 11 are screenshots of the real working environment outlier codes outputs for some particular surgery codes. So, the task of outlier detection observes the training data, and identifies the abnormal distribution of claim amount w.r.t to different surgeries which have taken place at various locations. The time complexity would increase if the data size increases but performance is comparatively good.

	A	B	C	D	E
1	Category	Surgery_C	Hospital_l	Claim_Amount	
2	M1	M1.1	RURAL	40000	
3	M1	M1.1	SEMI-URB	50000	
4	M1	M1.2	RURAL	40000	
5	M1	M1.2	SEMI-URB	80000	
6	M1	M1.2	URBAN	80000	
7	M1	M1.3	RURAL	45000	
8	M1	M1.3	SEMI-URB	60000	
9	M1	M1.3	URBAN	60000	
10	M1	M1.4	RURAL	75000	
11	M1	M1.4	SEMI-URB	130000	
12	M1	M1.4	URBAN	130000	
13	M1	M1.5	RURAL	6000	
14	M1	M1.5	SEMI-URB	115000	
15	M1	M1.5	URBAN	135000	
16	M1	M1.7	RURAL	85000	
17	M1	M1.7	SEMI-URB	80000	
18	M1	M1.7	URBAN	50000	
19	M1	M1.8	RURAL	60000	
20	M1	M1.8	SEMI-URB	60000	
21	M1	M1.8	URBAN	60000	
22	M10	M10.1	SEMI-URB	20000	
23	M10	M10.1	URBAN	32067	
24	M10	M10.1.2	URBAN	71250	
25	M10	M10.2	SEMI-URB	15000	
26	M10	M10.2	URBAN	30000	

Fig. 4. Screenshot of Transformed Feature subset by applying filter method and concept hierarchy on NTRSSV dataset.

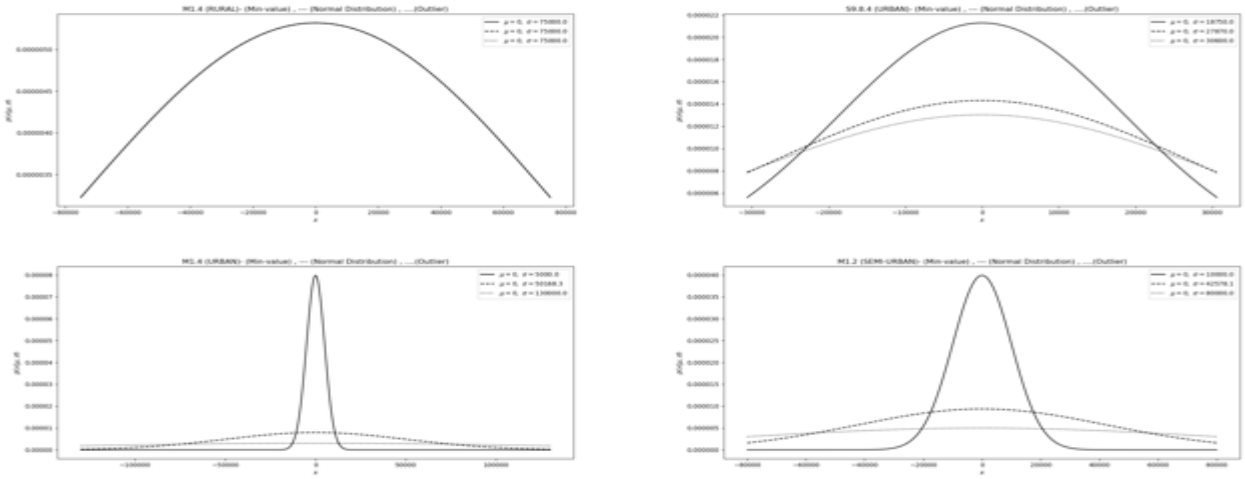

Fig. 5. PDF values of claims distribution for surgery codes M1.4_R, S9.8.4_U,M1.4_U,M10.3_U

Surgery_Code	Hospital_Location_CATEGORY	Claim amount
M1.1	RURAL	40000.000000
	SEMI-URBAN	30000.000000
M1.2	RURAL	40000.000000
	SEMI-URBAN	42578.126984
	URBAN	37692.307692
M1.3	RURAL	33333.333333
	SEMI-URBAN	32500.000000
	URBAN	28206.909091
M1.4	RURAL	75000.000000
	SEMI-URBAN	45879.310345
	URBAN	50168.307692
M1.5	RURAL	6000.000000
	SEMI-URBAN	43636.857143
	URBAN	54500.000000
M1.7	RURAL	41592.592593
	SEMI-URBAN	31931.407407
	URBAN	36021.300971
M1.8	RURAL	19750.000000
	SEMI-URBAN	28247.600000
	URBAN	35251.790698
M10.1	SEMI-URBAN	20000.000000
	URBAN	19783.130435
M10.1.2	URBAN	52000.000000
M10.2	SEMI-URBAN	15000.000000
	URBAN	19384.615385
M10.3	URBAN	23000.000000
M10.4	URBAN	8000.000000
M10.5	URBAN	18866.714286
M11.1.1	SEMI-URBAN	15946.970588
	URBAN	21510.000000
		...
S9.8.2	URBAN	39424.800000
S9.8.3	SEMI-URBAN	41866.000000
S9.8.4	RURAL	30000.000000

Fig. 6. Screenshot of list Initial observations of claim amount by calculating the average statitical mean of each sub category surgery code

```
M10
+--------------+-------+-------------+-------+
| Surgery_code | RURAL | SEMI-URBAN  | URBAN |
+==============+=======+=============+=======+
| M10.2        | -     | 15000       | -     |
+--------------+-------+-------------+-------+
---------------Exceeds Amount Max list---------------
+--------------+-------+-------------+-------+
| Surgery_code | RURAL | SEMI-URBAN  | URBAN |
+==============+=======+=============+=======+
| M10.2        | -     | 32067       | -     |
+--------------+-------+-------------+-------+
0
3
M11
+--------------+-------+-------------+-----------+
| Surgery_code | RURAL | SEMI-URBAN  |   URBAN   |
+==============+=======+=============+===========+
| M11.1.1      | -     | 15946.971   | -         |
+--------------+-------+-------------+-----------+
| M11.1.2      | -     | 16800       | 16333.333 |
+--------------+-------+-------------+-----------+
| M11.2.4      | -     | 14500       | -         |
+--------------+-------+-------------+-----------+
---------------Exceeds Amount Max list---------------
+--------------+-------+-------------+-------+
| Surgery_code | RURAL | SEMI-URBAN  | URBAN |
+==============+=======+=============+=======+
| M11.1.1      | -     | 71250       | -     |
+--------------+-------+-------------+-------+
| M11.1.2      | -     | 30000       | 25000 |
+--------------+-------+-------------+-------+
| M11.2.4      | -     | 32067       | -     |
+--------------+-------+-------------+-------+
1
```

Fig. 7. List of Outlier codes for sub surgery code M10, M11 w.r.t hos-loc-cat of category which exceeds amount of prior maximum cost

```
40000.0
1725
1725
2
_____ Oultiers in Dict format _____
['M1', 'M10', 'M11', 'M12', 'M2', 'M4', 'M5', 'M6', 'M7', 'M8', 'M9', 'S1', 'S10', 'S11', 'S12', 'S13', 'S14', 'S15', 'S16', 'S17', 'S2', 'S3', 'S4', 'S5', 'S6', 'S7', 'S8', 'S9']
[['M1.1', 'M1.2', 'M1.3', 'M1.4', 'M1.5', 'M1.7', 'M1.8'], ['M10.2'], ['M11.1.1', 'M11.1.2', 'M11.2.4'], ['M12.10', 'M12.12', 'M12.13', 'M12.15', 'M12.4', 'M12.5'], ['M2.1',
'M2.4', 'M2.5', 'M2.6', 'M2.7'], ['M4.1.1', 'M4.1.13', 'M4.1.14', 'M4.1.15', 'M4.1.16', 'M4.1.17', 'M4.1.18', 'M4.1.2', 'M4.1.3', 'M4.1.4', 'M4.1.5', 'M4.1.6', 'M4.1.7', 'M4.1.8',
'M4.2.1.1', 'M4.2.1.2', 'M4.2.1.3', 'M4.2.1.4', 'M4.2.1.6', 'M4.2.1.7', 'M4.2.2.2', 'M4.2.2.3', 'M4.2.3.1', 'M4.2.3.2', 'M4.2.3.3', 'M4.3.1.1', 'M4.3.1.2', 'M4.3.2.1', 'M4.3.2.2',
'M4.3.3.1', 'M4.3.3.4', 'M4.3.4.2', 'M4.3.4.3', 'M4.3.5.1', 'M4.3.5.2', 'M4.3.5.4', 'M4.3.6.1', 'M4.3.6.2', 'M4.3.6.3', 'M4.3.6.4'], ['M5.1.1', 'M5.1.2', 'M5.1.3', 'M5.6'],
['M6.1', 'M6.3', 'M6.4'], ['M7.1', 'M7.11', 'M7.3', 'M7.4', 'M7.5', 'M7.7', 'M7.8', 'M7.9'], ['M8.1', 'M8.2', 'M8.3', 'M8.4', 'M8.6', 'M8.7'], ['M9.1'], ['S1.1.1.13', 'S1.1.10',
'S1.1.11', 'S1.1.5.1', 'S1.1.5.5', 'S1.1.5.6', 'S1.1.6', 'S1.1.9', 'S1.2.1', 'S1.3.1.1', 'S1.3.1.10', 'S1.3.1.2', 'S1.3.1.3', 'S1.3.1.5', 'S1.3.1.6', 'S1.3.1.7', 'S1.3.1.8',
'S1.3.1.9', 'S1.3.2.1', 'S1.3.2.2', 'S1.3.4.8', 'S1.3.5.7', 'S1.3.5.8', 'S1.5.1', 'S1.5.2', 'S1.6.2', 'S1.7.1', 'S1.8.1'], ['S10.1.10', 'S10.1.12', 'S10.1.14', 'S10.1.18',
'S10.1.23', 'S10.1.24', 'S10.1.28', 'S10.1.6', 'S10.1.8', 'S10.2.1', 'S10.2.11', 'S10.2.5', 'S10.2.8', 'S10.3.3', 'S10.3.5', 'S10.4.3', 'S10.6.3'], ['S11.1.1', 'S11.1.10',
'S11.1.2', 'S11.1.3', 'S11.1.4', 'S11.1.5', 'S11.1.6', 'S11.1.9', 'S11.12.1', 'S11.14.1', 'S11.18.1', 'S11.18.2', 'S11.18.3', 'S11.2.2', 'S11.2.4', 'S11.2.5', 'S11.2.6', 'S11.22
.1', 'S11.24.1', 'S11.25.1', 'S11.25.3', 'S11.26.1', 'S11.27.1', 'S11.27.2', 'S11.28.6', 'S11.3.1', 'S11.3.10', 'S11.3.4', 'S11.3.5', 'S11.3.6', 'S11.3.8', 'S11.3.9', 'S11.31.3',
'S11.35.1', 'S11.35.2', 'S11.35.3', 'S11.35.5', 'S11.35.6', 'S11.35.8', 'S11.36.1', 'S11.4.2', 'S11.4.3', 'S11.4.4', 'S11.4.5', 'S11.5.1', 'S11.5.2', 'S11.5.3', 'S11.5.4',
'S11.6.1', 'S11.6.2', 'S11.6.3', 'S11.7.1', 'S11.7.2', 'S11.7.3'], ['S12.1.1', 'S12.1.2', 'S12.1.3', 'S12.1.4', 'S12.1.5', 'S12.1.7', 'S12.10.1', 'S12.10.2', 'S12.11.1', 'S12.11
.5', 'S12.12.1', 'S12.13.1', 'S12.13.2', 'S12.14.1', 'S12.14.2', 'S12.15.1', 'S12.16.1.1', 'S12.16.2.1', 'S12.17.1', 'S12.17.3', 'S12.2.1', 'S12.22.1', 'S12.24.1', 'S12.25.1',
'S12.27.1.2', 'S12.27.1.3', 'S12.28.1', 'S12.29.1', 'S12.3.1', 'S12.31.1', 'S12.32.1', 'S12.32.2', 'S12.4.1', 'S12.5.1', 'S12.6.1', 'S12.7.1.2', 'S12.7.2.1', 'S12.8.1', 'S12.9.1'],
['S13.1.1', 'S13.1.2', 'S13.1.3', 'S13.2.1', 'S13.2.3', 'S13.3.1.1', 'S13.3.1.2', 'S13.4.2.1'], ['S14.1.1', 'S14.1.2', 'S14.1.4', 'S14.12', 'S14.2.1.2', 'S14.2.1.3', 'S14.2.1.4',
'S14.2.2.1', 'S14.2.2.2', 'S14.2.2.3', 'S14.3.1', 'S14.4'], ['S15.1.2', 'S15.1.3', 'S15.2.1.1', 'S15.2.1.2', 'S15.3.1.1', 'S15.3.1.2', 'S15.4.1.1', 'S15.5.1', 'S15.7.1', 'S15.7
.2'], ['S2.3.3'], ['S3.2.3', 'S3.3.3', 'S3.3.4', 'S3.3.5', 'S3.6.2', 'S3.66'], ['S4.2.12', 'S4.2.3', 'S4.2.5', 'S4.2.9'], ['S5.1.1', 'S5.1.3', 'S5.1.4', 'S5.1.5', 'S5.2.1',
'S5.3.1', 'S5.4.1', 'S5.4.2', 'S5.4.3', 'S5.4.4', 'S5.4.5', 'S5.4.6', 'S5.5.1', 'S5.5.3', 'S5.6.1', 'S5.6.2', 'S5.7.1'], ['S6.11.3', 'S6.2.1', 'S6.2.4', 'S6.3.4', 'S6.5.1',
'S6.5.4', 'S6.5.6', 'S6.8.1'], ['S7.1.1.2', 'S7.1.1.3', 'S7.1.1.4', 'S7.1.4.1', 'S7.1.5.1', 'S7.1.5.2', 'S7.1.7.4', 'S7.11.1', 'S7.11.10', 'S7.11.13', 'S7.11.14', 'S7.2.1.2',
'S7.2.1.3', 'S7.2.11.3'], ['S8.1.2.5', 'S8.1.2.6', 'S8.1.2.8', 'S8.1.4.1', 'S8.1.4.2', 'S8.2.2', 'S8.2.6.1', 'S8.3.1', 'S8.3.5', 'S8.4.1', 'S8.4.3', 'S8.5.1', 'S8.5.7', 'S8.8.14'],
['S9.1.2', 'S9.10.1', 'S9.2.1', 'S9.2.3', 'S9.3.1', 'S9.3.2', 'S9.3.3', 'S9.3.4', 'S9.3.5', 'S9.3.6', 'S9.4.1.1', 'S9.4.1.2.1', 'S9.4.1.2.2', 'S9.4.3.1', 'S9.4.3.2.1', 'S9.4.3.2
.2', 'S9.4.4', 'S9.5.1', 'S9.6.1', 'S9.6.2', 'S9.6.3', 'S9.6.4', 'S9.6.8', 'S9.7.1', 'S9.7.2', 'S9.8.1', 'S9.8.4', 'S9.8.5', 'S9.8.6', 'S9.8.7', 'S9.9.1', 'S9.9.2', 'S9.9.3',
'S9.9.4', 'S9.9.7']]
7
M1
+--------------+-----------+-------------+-----------+
| Surgery_code |   RURAL   | SEMI-URBAN  |   URBAN   |
+==============+===========+=============+===========+
```

Fig. 8. Console output for list of outliers of surgery codes.

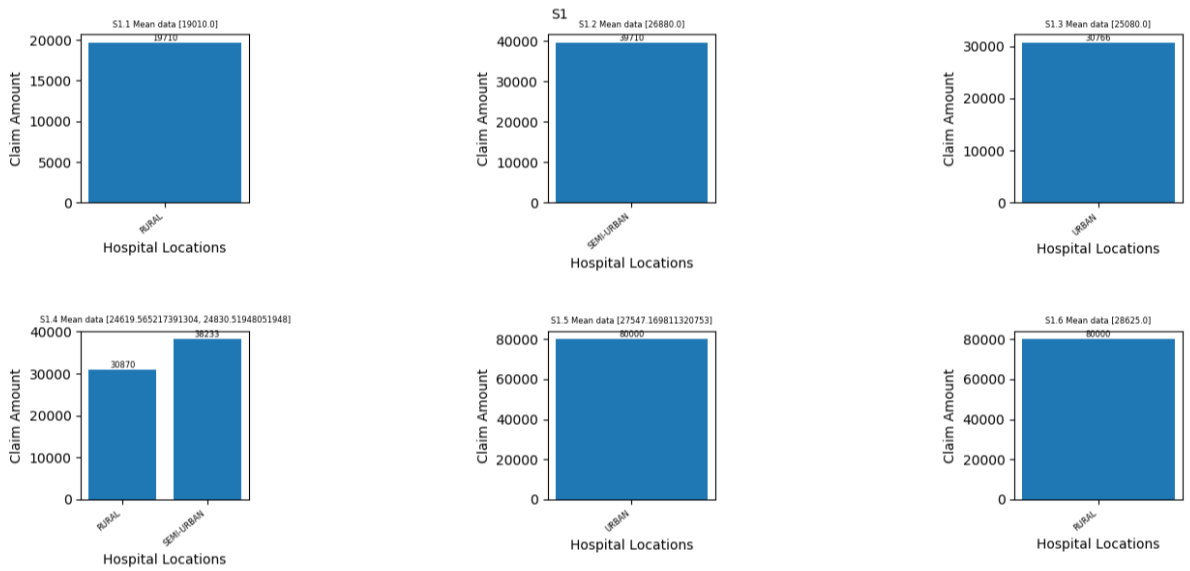

Fig. 9. Histogram-based Graph representation and outputs for some surgery codes. On x-axis shows the hospital location and y-axis represents the distribution of claim amount of the particular surgery code w.r.t hospital_loc_cat (rural,urban, semi-urban)

Pseudo code for Graph Representation (tit_g, sur_code, temp_loc_list,temp_outlier amount, normal amount_list):

1. For representing values in bar graph format
   fig axes=plt.subplots(2, 1)
   plt.suptitle(tit_g,fontsize=20)
   in axes 0 plot Normal Distribution
   in axes 1 plot Outliers data
   call the graph_representation function
   (tit_g,sur_code_data,
   loc_list_data,amount_list_data,mean_list_data)
   end
2. Represent the graphs x- axis as Hospital locations and Y-axis as claim amounts of all outliers and also show the normal distribution values for all surgery codes
3. Represent the graphs x- axis as Hospital locations and Y-axis as claim amounts of all outliers and also show the normal distribution values for all surgery codes.

Pseudo code for graph visualization using matplotlib in python for all categories of surgery codes and subcategories

M5

Average statistical mean value for the M5 sub category
Surgery code w.r.t Hospital loation category

Surgery_code	RURAL	SEMI-URBAN	URBAN
M5.1.1	9400.0	9744.68085106383	9239.47619047619
M5.1.2	23542.468013468013	23855.20887096774	24057.795281582952
M5.1.3	6000.0	18949.0	18225.535714285714
M5.6	-	15285.714285714286	-

Outliers cost for M5 sub category surgery codes
w.r.t Hospital location category

Surgery_code	RURAL	SEMI-URBAN	URBAN
M5.1.1	30000.0	25000.0	60000.0
M5.1.2	45000.0	60000.0	25000.0
M5.1.3	60000.0	30000.0	35000.0
M5.6	-	18000.0	-

Fig. 10. Screenshot for comparison of average statistical mean value and outlier cost for M5 subcategory surgery codes

Surgery_code	RURAL	SEMI-URBAN	URBAN
S5.1.1	45900.0	25000.0	-
S5.1.3	25500.0	115700.0	118014.0
S5.1.4	95000	95000.0	96900.0
S5.1.5	96900.0	-	-
S5.2.1	25000.0	32746.0	32746.0
S5.3.1	32746.0	-	-
S5.4.1	30000.0	44540.0	44540.0
S5.4.2	20000.0	81010.0	21400.0
S5.4.3	30000.0	46790.0	46790.0
S5.4.4	58850.0	37450.0	37450.0
S5.4.5	32950.0	32950.0	80000.0
S5.4.6	60000.0	60000.0	60000.0
S5.5.1	-	65000.0	65000.0
S5.5.3	100000.0	96000.0	-
S5.6.1	-	80000.0	-
S5.6.2	-	115260.0	-

Fig. 11. Screen shot of the results for Outlier codes of few surgery codes with respect to hospital location category

The results in the console environment and table representation do have some lack of quick understanding of results.The following observation and analysis are drawn with help of matplotlib in Python.

5. Conclusion

With the experimentation of outlier detection in medical claims tends to be proactive for finding the suspicious (fraud) activities. This model works for the medical insurance organization to identify the medical fraud w.r.t. to claiming behavior. The recommended methodology uses distance-based and statistical mean measure to the find degree of outlier object which deviates from the normal object from its neighborhood objects. The distance-based outlier codes analyzed the properties of the objects or instances with Gaussian probability density function and identify the false distribution of instances. As far the limitation it undergone testing with single dataset. The scalability of the proposed methodology is high and computational cost is average. Based on the real world applications it can be extended to high dimensional data also, but the efficiency in terms of time complexity (execution) goes low. As part of upcoming work, to improvise accuracy and to test with more datasets. Further, to develop a decision support system with help of classification algorithm for providing decision on Medicare data.

References

1. Varun Chandola, Arindam Banerjee, and Vipin kumar, Outlier detection survey, ACM Computing, (2007).
2. Tan P.N, Steinbach M., and Kumar V. Introduction to Data Mining. Addison-Wesley, Chapter 2, pp.19-96, (2005a).
3. Mitchell T. M, Machine Learning. McGraw-Hill Higher Education (1997).
4. Vapnik, V. N. The nature of statistical learning theory. Springer-Verlag New York, Inc., New York, NY, USA, (1995).
5. Abe, N., Zadrozny, B., and Langford, J., Proc. 12th ACM SIGKDD on Knowledge Discovery and data mining, New York, Outlier detection by active learning. NY, USA, pp.504-509, (2006).
6. Dasgupta, D. and Majumdar, N., Proc. IEEE Conference on Evolutionary Computation. Hawaii, Anomaly detection in multidimensional data using negative selection algorithm, pp.1039-1044, (2002).
7. Markou, M. and Singh, S., Novelty detection: a review-part 1: Statistical approaches Signal Processing 83, 12, pp.2481-2497, (2003a).
8. Markou, M. and Singh, S., Novelty detection: a review-part 2: Neural network based approaches. Signal Processing 83, 12, pp.2499-2521, (2003b).
9. Kamaljeet Kaur and Atul Garg, Comparative study of outlier detection algorithms, International journal of computer applications, Vol.147, No.9, pp.21-25, (2016).
10. Laurikkala, J., Juhola1, M., and Kentala, E. Informal identification of outliers medical data. In Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology, pp.20-24,(2000).
11. Ghosh, S. and Reilly, D. L. Proc. 27th Annual Hawaii International Conference on System Science, Los Alamitos, Credit card fraud detection with a neural network, (1994).
12. He Z., Xu, X., and Deng, S. Discovering cluster-based local outliers. Pattern recognition letters 24, 9-10, pp.1641-1650, (2003).
13. R. J. Bolton and J. H. David. Statistical fraud detection: A review. Statistical Science, vol.17, (2002).
14. E. Elmer. Pro_ling Machines: Mapping the Personal Information Economy, volume 1. MIT Press, (2004).
15. Guido Cornelis van capelleveen, B.Sc. Thesis, Utrecht University, December, (2013).
16. Markus Goldstein and Andreas Bengel, Histogram-based Outlier Score (HBOS): A fast unsupervised Anomaly Detection, German Research Center for Artificial Intelligence (DFKI), (2012).
17. Xi Jingke. Proc. 2nd International Symposium IEEE, Outlier detection algorithms in data mining. Intelligent Information Technology Application, (2008).
18. Indukuri BhaskaraRaju, PhD Thesis on Healthcare utilization to Socio-Economic Status in Rural and Urban Areas of Andhra Pradesh , BITS, PILANI, Rajasthan, (2015).
19. Dallas Thornton, Guido van Capelleveen, Mannes Poel, Proc. Int conference on Enterprise Information Systems, Outlier-based Health insurance Fraud Detection for U.S. Medicaid data, pp.684-694,(2014).
20. Milou Meltzer, MSC Thesis, Outlier detection in datasets with a mixed attribute,Vrije Universiteit Amsterdam, (2015).
21. JianWu, Runtong Zhang, Xiaopu Shang and FuzhiChu, Proc. 2ndInt.Conference on Artificial Intelligence and Engineering Applications, Medical Insurance Fraud Recognition Based on Improved Outlier Detection Algorithm, pp.765-772,(2017).
22 Christy.A, Meera Gandhi.G, S.Vaithyasubramanian, Proc. Int.Symposium on Big Data and Cloud Computing, Cluster-Based Outlier Detection Algorithm for Healthcare data, pp.209-215,(2015).

23. Zhang, Weijia, and Xiaofeng He. Proc. Int. Conf IEEE, An Anomaly Detection Method for Medicare Fraud Detection. In Big Knowledge (ICBK), pp. 309-314. IEEE, (2017).
24. Chandola Varun, Sreenivas R. Sukumar, and Jack C. Schryver. In Proceedings of the 19th ACM SIGKDD Knowledge discovery from massive healthcare claims data. pp. 1312-1320, (2013).
25. Archana Kadam and Sager G Powar, Hybrid approach to outlier detection in medical data set, Asian Journal of computer science and Technology, Vol.6, No.2, pp.18-22, (2017).
26. Yijun Lu, PhDThesis on Concept hierarchy in data mining: Specification Generation and Implementation, Simon Fraser University, December (1997).
27. Xuanfan Wu, Metrics, Techniques, and tools of Anomaly detection: A Survey, ebook,
pp. 1-12.
28. Travaille, Roland M Muller, Dallas Thornton, and Jos Van Hillegersberg, Electronic fraud detection in U.S. Medicaid healthcare program: Lessons Learned from other industries, Proc.7th Americans conference on information systems, pp.1-10, (2011).
29. Tang, MingJian, et al. Proc. of the Ninth Australasian Data Mining Conference, Unsupervised fraud detection in Medicare Australia. Volume 121, (2011).
30. Tsoi, A.C., Zhang, S., and Hagenbuchner, M., IEEE Transactions on Knowledge & Data Engineering, Pattern discovery on Australian medical claims data-a systematic approach, pp.1420-1435, (2005).
31. http://www.ntrvaidyaseva.ap.gov.in/web/guest/explore_data.