ISSN:0975-9646

Outlier Detection for Multidimensional Medical Data

Anbarasi.M.S
Assistant Professor, Dept. of Information Technology
Pondicherry Engineering College,
Puducherry, India

Ghaayathri.S, Kamaleswari.R, Abirami.I
B.Tech(IT) Final year, Pondicherry Engineering College,
Puducherry,India

***Abstract*— The knowledge-rich nature of the Medical Information domain has made it an ideal environment where knowledge on data mining should have to be unearthed from large data collection for dialysis' of growing unknown diseases. Outlier detection is an important research problem that aims to find objects that are considerably dissimilar, exceptional and inconsistent in the database. Medical application is a high dimensional domain hence determining outliers is found to be very tedious due to curse of dimensionality. Most of the existing outlier detection methods detect the so-called point outliers from vector-like data sets. In this paper, clustering technique is used to cluster for multidimensional data. The draw back in clustering is overcome by auto K-generation as first process. Then the outliers are deducted by Thompson's Tau method which is further enhanced by max-flow min-cut theorem to find the uniqueness of outliers in multidimensional Medical data.**

Keywords— Outlier detection, Multi dimensional data, Outliers, Clustering, Min cut theorem, k means.

I.INTRODUCTION

Due to incredible growth of medical dataset, conventional data base querying methods are inadequate to extract useful information, so researches nowadays are focussed to develop new techniques to meet the raised requirements. The increase in dimensionality of data gives rise to a number of new computational challenges not only due to the increase in number of data objects but also due to the increase in number of attributes.

There are various origins of outliers. With the growth of the medical dataset day by day, the process of determining outliers becomes more complex and tedious. K-means is a well known partitioning based clustering technique that attempts to find a user specified number of clusters represented by their centroids. In this system, the enhanced auto K- generation for K-means is done as the first step. But its output is quite sensitive to initial positions of cluster centers. Again, the number of distance calculations increases exponentially with the increase of the dimensionality of the data. As quality of the final clusters heavily depends on the selection of the initial centroids, a new method is used to choose such data objects as initial centroids. After clustering left out points are deducted by Thompson's Tau method. These points are determined and checked out if they can be accommodated in any of the cluster by max flow min-cut theorem. As a result, the number of outliers is minimised and uniqueness in outliers are deducted as a goal of this work.

II.RELATED WORK

(Sheng-Yi Jiang and Ai-Min Yang, 2009) [1] Have proposed a system which determines outliers in two stages. In first stage the dataset is clustered whereas in the second stage the gained clusters are categorised as "normal cluster" or "outlier cluster" according to an outlier factor, and finally confirm the outlier objects.

(K. A. Abdul Nazeer and M. P. Sebastian,2009) [2] have proposed an efficient algorithm which determines the initial centroids rather than assigning arbitrary centriods. (Robert Tibshirani et al., 2003 [4]) have proposed a method to automatically generate the number of clusters as input to the k means clustering algorithm.

(Jani Posio et al., 2008) [3] Have proposed a method for determining outliers in 2D temperature data.

III.PROPOSED WORK

The input to the proposed system is multi dimensional medical data and the output is simulated using matlab. The three modules are clustering, outlier detection and minimising outlier.

The architecture of the proposed system is shown in figure 1.1

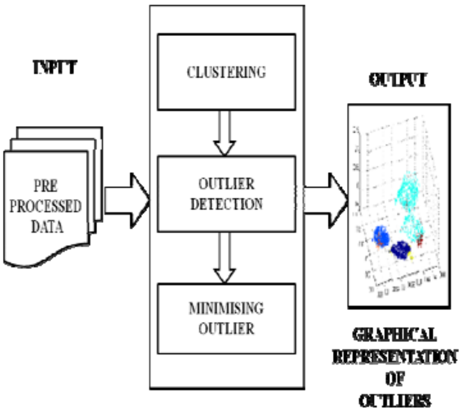


Fig: 1 Architecture diagram

The various modules to be discussed are:

A.Clustering

Cluster analysis is one of the major data analysis methods widely used for many practical applications in emerging areas. Clustering is the process of finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups as shown below:

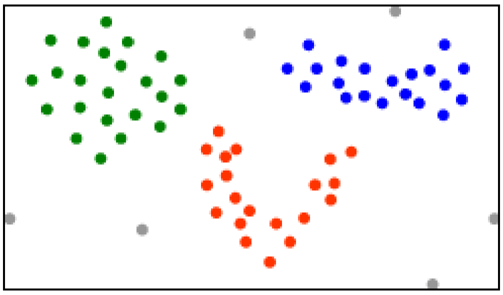


Fig 2 clustering

K-means [10], [9] is a commonly used partitioning based clustering technique that tries to find a user specified number of clusters. Although K-means is simple and can be used for a wide variety of data types, it is quite sensitive to initial positions of cluster centers.The k-means method is simple and fast, that works as follows:

- Arbitrarily choose k initial seeds
- Assign each object to the group that has the closest centroid
- Recalculate the positions of the centroids
- Repeat steps 2 and 3 until the positions of the centroids no longer changes

The flow chart for proposed k means algorithm is shown below:

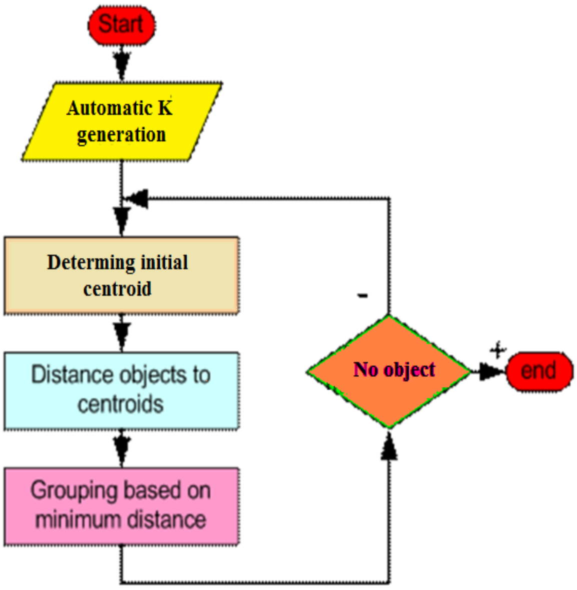

Start
Automatic K generation
Determing initial centroid
Distance objects to centroids
Grouping based on minimum distance
No object
-
+
end

Fig 3 enhanced k means clustering

B. Outlier Detection

An outlier [5], [6] is defined as a data point which is very different from the rest of the data based on some measure. The output of the clustering module is given as input to the outlier detection module to determine the outliers. The outliers are detected using the Thompson's Tau method which determines outliers.

Dimensionality deduction techniques, such as Principal Component Analysis (PCA) [7], [8] can be applied to the medical data before outlier detection is performed. In PCA, the medical data is projected onto lower dimensional data. PCA is used to perform feature selection and can be considered as the pre-processing work for outlier detection.

The modified Thompson tau technique is a statistical method for deciding whether to keep or discard suspected outliers in a sample of a single variable. Here is the procedure:

- The sample mean x and the sample standard deviation S are calculated in the usual fashion.
- For each data point, the absolute value of the deviation is calculated

$$\delta_{i=} |d_i|=|x_i—\bar{x}|$$

- The data point most suspected as a possible outlier is the data point with the maximum value of δ_i.
- The value of the modified Thompson τ (Greek letter tau) is calculated from the critical value of the student's t PDF, and is therefore a function of the number of data points n in the sample.
  τ is obtained from the expression

$$tau= (t_{\alpha/2}.(n-1))/(\sqrt{n}.\sqrt{(n-2+t_{\alpha/2})})$$

Where

- n is the number of data points
- tα/2 is the critical student's t value, based on α = 0.05 and df = n-2 (note that here df = n-2 instead of n-1). In Excel, we calculate tα/2 as TINV(α, df), i.e., here tα/2 = TINV(α, n-2)
  The above graph shows an outlier detected using Thompson's Tau.

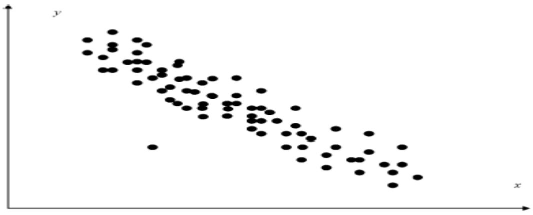


Fig 4 outlier detection

The flow chart for Thompson tau method is shown:

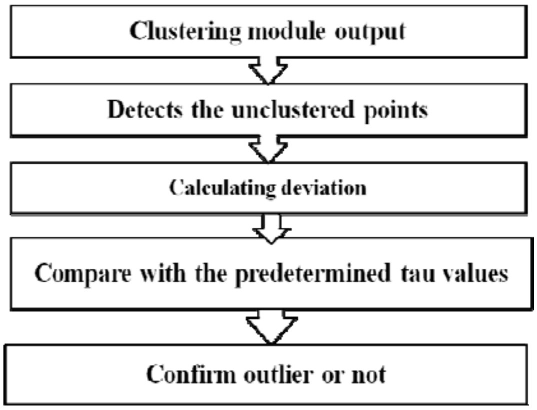
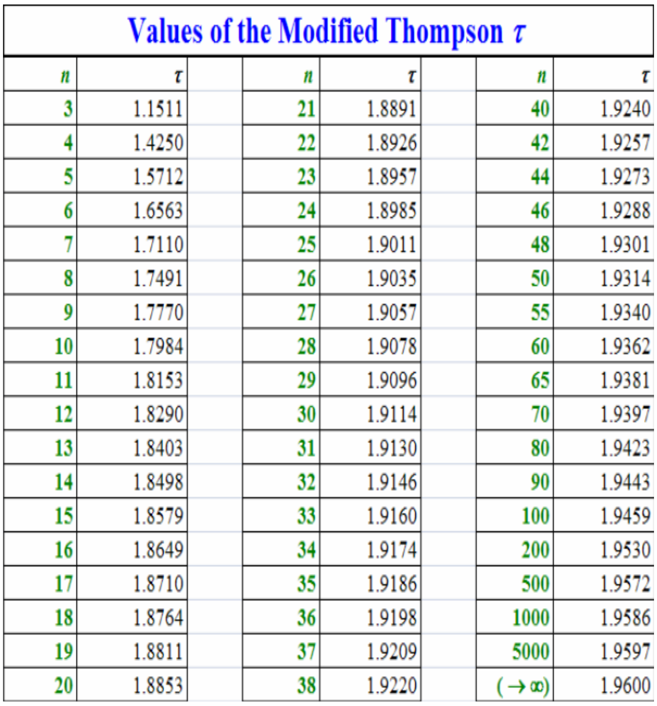

Clustering module output
Detects the unclustered points
Calculating deviation
Compare with the predetermined tau values
Confirm outlier or not

Fig 5 Flow chart for Thompson's Tau method

From the figure 1.3, it is evident that the unclustered points are detected as outliers and using Thompson's Tau method, the outliers are confirmed using the table shown below..

A table of the modified Thompson τ is provided below:



Values of the Modified Thompson τ

n	τ	n	τ	n	τ
3	1.1511	21	1.8891	40	1.9240
4	1.4250	22	1.8926	42	1.9257
5	1.5712	23	1.8957	44	1.9273
6	1.6563	24	1.8985	46	1.9288
7	1.7110	25	1.9011	48	1.9301
8	1.7491	26	1.9035	50	1.9314
9	1.7770	27	1.9057	55	1.9340
10	1.7984	28	1.9078	60	1.9362
11	1.8153	29	1.9096	65	1.9381
12	1.8290	30	1.9114	70	1.9397
13	1.8403	31	1.9130	80	1.9423
14	1.8498	32	1.9146	90	1.9443
15	1.8579	33	1.9160	100	1.9459
16	1.8649	34	1.9174	200	1.9530
17	1.8710	35	1.9186	500	1.9572
18	1.8764	36	1.9198	1000	1.9586
19	1.8811	37	1.9209	5000	1.9597
20	1.8853	38	1.9220	(→∞)	1.9600

Fig: 6 Modified Thompson Tau table

The Thompson's Tau method is an existing method using we are going to detect outliers and finally output of this algorithm is fed as input to the maxflow/mincut algorithm as a result of which the unique outliers in multi dimensional medical data is determined.

C. Minimising Outliers

As a result of k means clustering, there will be some points left out which cannot be included in any of clusters due to their unique properties. The distances between these outliers and existing clusters are determined.

The min cut theorem is used for reducing the outlier cluster distances and accommodating these outliers in the approximate clusters such the number of outliers can be reduced.

As a result of K-means clustering, there will be some points left out which cannot be included in any of clusters due to their unique properties. The distances between these outliers and existing clusters are determined.

The Min cut theorem is used for reducing the outlier cluster distances and accommodating these outliers in the approximate clusters such the number of outliers can be reduced.

1) Masking Effect

It is said that one outlier masks a second outlier, if the second outlier can be considered as an outlier only by itself, but not in the presence of the first outlier. Thus, after the deletion of the first outlier the second instance is emerged as an outlier.

2) Maxflow/ Mincut

A common scenario is to use a graph to represent a "flow network" and use it to answer questions about material flows

- Flow is the rate that material moves through the network
- Each directed edge is a conduit for the material with some stated capacity
- Vertices are connection points but do not collect material flow into a vertex must equal the flow leaving the vertex, flow conservation.

3) Maxflow

The maximum flow problem refers to finding the most suitable & feasible way through a single sourced & sinks network. It is also seen as the maximum amount of flow that can achieve from source to destination which is an incredibly important consideration especially in data networks where maximum throughput and minimum delay are preferred. Finding the maximum flow involves looking at all the possible routes between source and destination. Maximum flow can be found by assigning flow to each link in the network so that the total flow from source to destination is as large as possible. The maximum source to sink flow in a network is equal to the minimum source to sink cut in the network, which makes up the Maximum Flow minimum cut theorem.

4) Mincut

A cut is any set of directed links containing at least one link in every path from origin node to destination node. This means if the links in the cut are removed the flow from the origin to destination is completely cut off. The cut value is the sum of the flow capacities in the origin to node direction over all the links. The minimum cut problem is to find the cut across the network that has the minimum cut value over all possible cuts. The maximum flow problem is closely related to the minimum cut problem, creating the maximum flow minimum cut theorem.

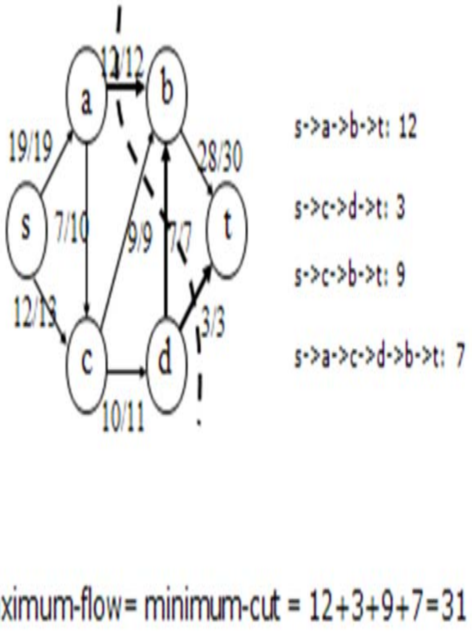


Fig 7 Mincut/Max flow

In the figure 1.6, an example of how the maxflow/mincut theorem can be applied is shown.

5) Maxflow/Mincut Algorithm

Step1: find the flow from source to sink, get the cut separating s and t, and use the smaller side as the candidate outlier or outlier group.

Step2: remove the candidate outlier or outlier groups from the graph.

Step3: select the next source; go back to 3 until the stop criterion.

Step4: adjusting the graph and adjust the maximum flow.

Most of the conventional outlier detection techniques are only applicable to relatively low dimensional static. Because they use the full set of attributes for outlier detection, thus they are not able to detect projected outliers. Recently, there is some emerging work in dealing with outlier detection either in high-dimensional static data.

IV.RESULT AND DISCUSSION

It encounters difficulties to identify outliers if data is not uniformly distributed. As shown in Fig: 1.7 C_1 contains 400 loosely distributed points, C_2 has 100 tightly condensed points, 2 outlier points o_1, and o2 as an outlier

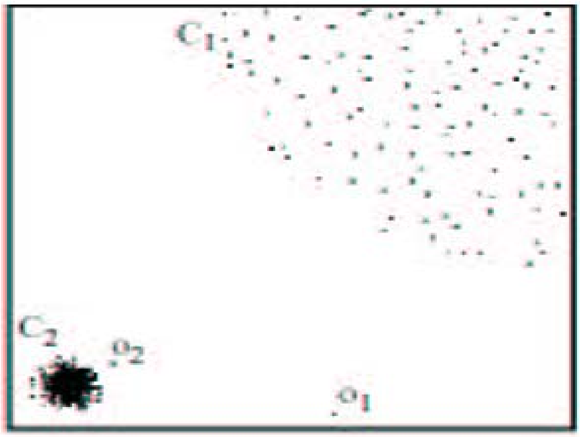


Fig: 1.8 Outlier deducted

V.CONCLUSION

Finding outliers is an important task in data mining. Outlier detection as a branch of data mining has many important applications and deserves more attention from data mining community. We conclude from our review of existing outlier detection schemes and clustering methods that they all suffer from the fact that they either depend on pre-specified values for the scale parameters or the fraction of inliers. These two main issues make them very sensitive to initialization; or they have to form a quasi exhaustive search on these parameters, which makes them require a very high computational cost.

VI.REFERENCES

[1] Sheng-Yi Jiang and Ai-Min Yang, “Framework of Clustering-Based Outlier Detection”, Sixth International IEEE Conference on Fuzzy Systems and Knowledge Discovery, ISBN: 978-0-7695-3735-1/09, 2009.

[2] K. A. Abdul Nazeer and M. P. Sebastian, “Improving the Accuracy and Efficiency of the k-means Clustering Algorithm”, Proceedings of the World Congress on Engineering, 2009.

[3] Jani Posio, Kauko Leiviskä, Jari Ruuska and Paavo Ruha, “Outlier Detection for 2D Temperature Data”, the International Federation of Automatic Control, 2008.

[4] Robert Tibshirani, Guenther Valther and Triver Hasthi, "Estimating number of clusters in a data set via Gap statistics", Royal Society, 2003.

[5] Charu C. Aggarwal and Philip S. Yu, "Outlier Detection for High Dimensional Data".

[6] Mohamed Bouguessa and Shengrui Wang, "Mining Projected Clusters in High-Dimensional Spaces", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 4, APRIL 2009.

[7] Anna Koufakou and Michael Georgiopoulos A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes, Data Mining and Knowledge Discovery (2010) 20:259–289, DOI 10.1007/s10618-009-0148-z.

[8] Rajashree Dash, Debahuti Mishra, Amiya Kumar Rath and Milu Acharya, "A hybridized K-means clustering approach for high dimensional dataset", International Journal of Engineering, Science and Technology Vol. 2, No. 2, 2010, pp. 59-66.

[9] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman and Angela Y. Wu, An Efficient k-Means Clustering Algorithm: Analysis and Implementation, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 7, JULY 2002.

[10] K. Karteeka Pavan, Allam Appa Rao, A.V. Dattatreya Rao and G.R. Sridhar, Single Pass Seed Selection Algorithm for k-Means, Journal of Computer Science 6 (1) ISSN 1549-3636 , pp: 60-66, 2010.