ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

Outlier Detection using Kmeans and Neural Network in Data Mining

Parmeet kaur
Department of computer science
Punjab Technical University, Jalandhar, India

***Abstract* - Outlier detection has been used to detect the outlier and, where appropriate, eliminate outliers from various types of data. It has vital applications in the field of fraud detection, network robustness analysis, Insider Trading Detection, email spam detection, Medical and Public Health Outlier Detection, Industrial Damage Detection, Image processing fraud detection, marketing, network sensors and intrusion detection. In this paper, we propose a kmean clustering and neural network as novel to detect the outlier in network analysis. Especially in a social network, k means clustering and neural network is used to find the community overlapped user in the network as well as it finds more kclique which describe the strong coupling of data. In this paper, we propose that this method is efficient to find out outlier in social network analyses. Moreover, we show the effectiveness of this new method using the experiments data. (Abstract)**

Index Terms: Outlier Detection; Network Data; Adjacency Matrix; Kmeans Clustering; Neural Network. (keywords)

1. INTRODUCTION

Data mining [1] is a procedure of extracting useful information and ultimately understandable information from huge datasets and then using it for organization's decision making process [6]. Still, there huge problems exist in mining datasets such as incomplete data, incorrect results, duplicity of data, the value of attributes is unspecific and outlier. Outlier detection is an essential task of data mining that is mainly focused on the discovery of items that are exceptional when contrasted with a group of observations that are measured typical. Outlier is a data item that does not match to the normal points characterizing the data set. Finding anomalous items among the data items is the basic idea to detect an outlier. Considerable research has been done in outlier detection and these are divided into different types with respect to the detection approach being used. These techniques include Cluster based methods [2], Classification based methods, Distance base method [3], Nearest Neighbor based methods [4], linear method [5] and Statistical based methods. In the Cluster-based approach, groups of homogenous types of items are formed. Cluster analysis refers to formulate the group of items that are more related to each other and others which is different from the items in other cluster. In the Classification-based approach [3] a model is generated from a group of data items with labels and then a test item is classified into one of the classes using appropriate testing. Nearest Neighbor based methods [4] involve similarity distance or distance measures; which are defined between data items. A statistical method is a mathematical based model, in which mathematical equations relate to one or more items and possibly other non-random items. In this paper, we discuss a new method to find out an outlier that is based on a graph dataset. This method reduces the search space efficiently and takes less time to find out outlier.

In this paper, we propose an approach to detect outlier from network data using Kmean clustering and neural network. Conceptually, we define the outlier as the influential user whose performance is higher than other users. The remainder of this paper is organized as follows. In section II terms and definitions are discussed, approach of proposed algorithm and proposed methodology with algorithm is given in section III, and conclusion is described in IV.

There are various types of algorithm exist to detect the outlier in data mining.

Marghny and Taloba (2011) the genetic algorithm is used in order to detect the outlier within the complex networks. The genetic algorithm utilizes the iterative approach in order to generate optimal result. There exist phases of GA such as selection, genetic operation, mutation and crossover. The genetic algorithm randomly varies the population in order to check the node which does not satisfy the properties of the group. This will give the outlier within the complex network [7].

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

Shashikla, George, and Shujaee (2015) the outlier detection is conducted by the use of proposed system. The proposed system handles the betweennees centrality (BEC) in order to determine the points which lie inside the cluster and the point which does not belong to the community. The community overlapping with outlier elimination is suggested through the proposed approach. The rate is considerable enhanced by the use of proposed approach [8].

Kumar, Kumar, and Singh (2013) the clustering based approach is used for the purpose of outlier detection. The outlier is the node which does not belong to the group. The attributes have to be utilized in order to detect the same. The attributes if not satisfied then the node is declared as outlier and then it is eliminated from the group. [9]

Palla, Gergely, ImreDerényi, IllésFarkas, and TamásVicsek(2005)described the first KClique based clique percolation method. The clique percolation method builds up the communities from k-cliques, which keep up a correspondence to complete (fully connected) sub-graphs of k nodes. Two k-cliques are considered adjacent if cliques share k − 1 nodes. A community is defined as the maximal union of k-cliques communities that can be reached from each other through a sequence of adjacent k-cliques.in this approach, first of all, find the maximal clique of a graph then find others clique .then create clique overlapped matrix to detect the overlapped community and set threshold matrix value is equal to k-1.in this matrix, if the value in row and column diagonal is more than k-1 then convert into one, otherwise zero. In clique matrix, 1 means overlapped and 0 means no overlapped nodes.[10]

Yildiz, Hakan, and Christopher Kruegel (2012) described a clique based algorithm to detect the overlapped community. KClique is a clique or subgraph of the graph with k nodes and in KClique and the KClique community is a union of all k cliques that can be reached from each other through a sequence of adjacent KClique. Two cliques are adjacent to each other if they share k-1 nodes with each other. In KClique method, first of all, find all the clique or communities with KClique of size k. after finding all the cliques. In last, union of all the nodes of KClique to detect overlapped communities. [11]

II. TERMS AND DEFINITIONS

•***Adjacency Matrix:*** is also called connection matrix and it shows the connection between rows v(i) and columns v(j) of a graph datasets v(i,j). If the two nodes are connected to each other than it writes 1 in Adjacency Matrix, otherwise 0 for no connection.

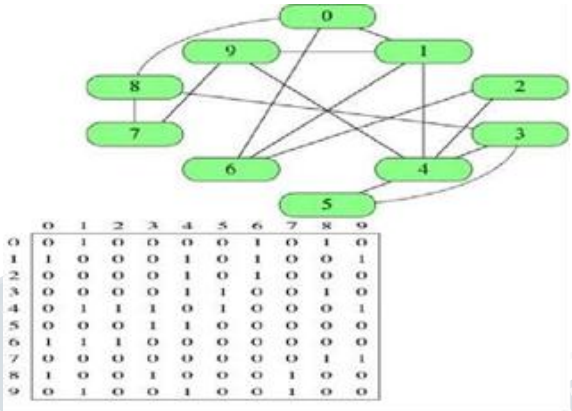
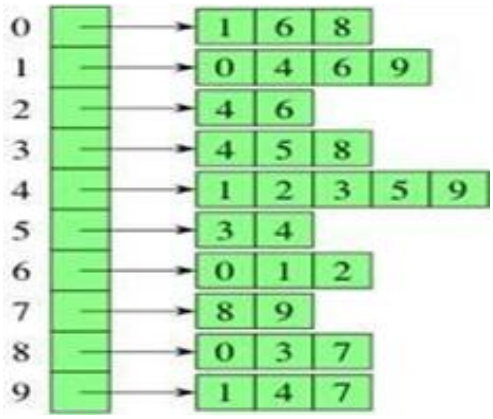


Figure 1 shows the network and adjacency matrix for undirected graph. 1, 2, 3, 4, 5, 6,7,8,9 represent the nodes. In the diagonal, all values are zero and if two nodes are connected, the matrix is denoted by the value of 1, otherwise 0.

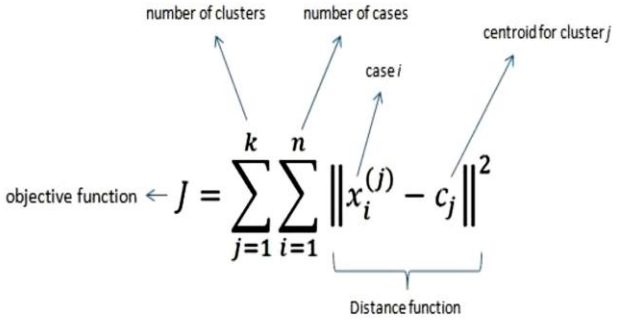
•***Adjacency List:*** it combines the adjacency matrix with edge. We represent the Adjacency List in array format. Here's an adjacency-list representation of above social network graph.



ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

•***Kmeans clustering:*** K-means clustering refers to partitioning n-observations into k clusters in which each observation belongs to one cluster with the nearest mean (centroid).



$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Working of Kmean clustering algoritgm-

1. Clusters the dataset into k groups where k is predefined.
2. Select k points at random as centers of cluster.
3. Assign objects to their closest cluster center according to the Euclidean distance parameter function.
4. Calculate the mean of all objects in each cluster.
5. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster (in successive round).

•***Neural Network:*** Neural Network algorithm is used to compute the maximum member function. This neural network algorithm will provide optimal value or influential user.

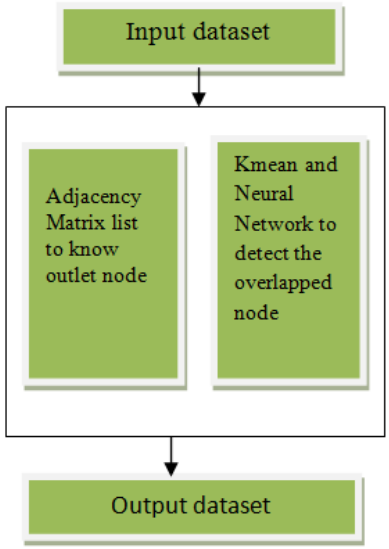
The proposed outlier detection method is based on Kmean clustering and fuzzy min max neural network for network data. For each node, we calculate the adjacency link with adjacency matrix. This is described in figure 1.

If a node has adjacency link is equal to zero, then we eliminate that node because our main focus to detect optimal user from network data. After eliminating the outlet node, then apply Kmeans clustering to make the clusters with Euclidian distance. In last to detect the optimal user apply fuzzy min max neural network which will provide overlapped user or optimal user. It takes less space to find the optimal user because we eliminate the outlet node, otherwise outlets will take some time and space to compute the result.

As our proposed method needs to find the adjacency matrix for each node, it is require detecting the outlet.

III. PROPOSED METHODOLOGY

The proposed methodology uses two methods to detect the outliers, one for eliminating the outlet node, second for outlier detection with kmean clustering and fuzzy min max neural network. In this method, firstly detect the outlets node with k clique method with help of adjacency matrix of network data. Main focus is on outlier detection with Kmean and neural network techniques and methods, which are used to detect the outlier from huge amount of data.



System architecture of work and proposed methodology step by step is given below:

A. System divided into three phases:

Phase 1: first of all, take a database which contains number of nodes. Then make adjacency matrix list of that dataset. If a node has adjacency less than 1 in adjacency list, then they will be declaring as outlet node.

Phase 2: Kmean is a method of clustering. It divides the data into k numbers of clusters

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

Phase 3: the output of Kmean is given to fuzzy min max neural network as input. After that neural network generate the output dataset.

1. Dataset:
Here we used machine learning dataset which is provided by Stanford University. This dataset contains number of node and connection of nodes with others nodes.

B. Proposed algorithm:
The proposed algorithm will take the adjacency matrix from the graph and then eliminate the nodes having 0 in the corresponding row matrix. Thus time will be saved. The Kclique algorithm will be modified in this case. The proposed algorithm will be described as follows.

Algorithm: Given as input a simple graph G with n vertices marked 1, 2… n, search for a clique of size at least k. At each stage, if the clique obtained has size at least k, and then stop.

KNOD (Graph G, Node n)

a) Obtain Adjacency(A)=Adj(G)
b) Set i=0
c) Repeat while i<=n
d) Check Adj_i
e) If(Adj_i>0)
f) Accept the node(AC_i)=N_i
g) Else
h) Reject the node
i) End of if
j) Move to the next node
k) I=i+1
l) End of loop
m) Perform K-Means to identify distinct cluster
n) Calculate optimal values using Fuzzy Min Max Technique

Tables 1 describe the comparison of Kclique (considering outlet node) with proposed method in which we don't consider outlet. The shows that proposed method is efficient on large number of datasets and it takes less time to detect the outlier

parameter	Kclique	Proposed Method
Number of Nodes	143	143
Cliques	3	8
Time Consumed	8.05ms	2.13ms
modularity	0.7686	0.82

Table 1: Comparison of KClique and Proposed Method

IV. CONCLUSION

The data mining is the mechanism which is used in order to filter the information which is extracted. The tools of data mining are present. The tool which is used in the existing system is WEKA. The proposed utilizes MATLAB for the purpose of data mining. The data mining is used to extract the information from the large dataset. The dataset is derived from the machine learning UCI website. The dataset generates the graph which is used to generate the information regarding the complex network. The information is represented graphically enhancing the working the existing system.
The community overlapping is the mechanism by which nodes with more than one interception are detected and avoided. The outlier consumes more bandwidth as compared to normal situation where outliers are absent. The complex network is used in the proposed system to detect the outlier in the existing system. The complex network is represented with the help of graph. The dataset is derived from the UCI website. The machine learning datasets are derived from that website. The speed of the existing system is reduced since nodes with 0 degree is also consider. When the control reaches the outlier node it had to shift backward which waste time.
The proposed system eliminates the outlier considerations by considering only those nodes in the complex networks which has degree more than 1. This way only fair node comes into picture. The time consumption is reduced by the way of proposed technique. The proposed system also determines the total number of community overlapping nodes by the way of fuzzy system. The fuzzy system is considered by the use of fuzzy in MATLAB. The fuzzy system describes the rules whose result is either true or false. The result if true the node is considered otherwise

node is rejected. This way outlier node is eliminated from the simulation.
Finally the performance of the proposed system is analyzed in terms of the bar graph. The bar graph shows that the proposed system result in terms of the time is better as compared to the existing system. The number of cliques detected is also enhanced by the use of proposed system.

FUTURE WORK

Overlapping community detection is still a challenge. Though there are several proposed methods, but most of them take a huge amount of processing time. So emphasis should be given to effective algorithms which will be able to detect communities in a huge social network in allowable time. In these work only unweighted and undirected networks has been taken into consideration. In future weighted and directed networks are needed to be considered for community detection. Now days almost all social networks are dynamic that are some members are joining and some are leaving every moment. So it will be great if communities can be detected in dynamic networks.

REFERENCES

[1] Lekhi, N., & Mahajan, M. (2015). Outlier Reduction using Hybrid Approach in Data Mining. International Journal of Modern Education and Computer Science, 7(5), 43.

[2] Pamula, R., Deka, J. K., & Nandi, S. (2011, February). An outlier detection method based on clustering. In Emerging Applications of Information Technology (EAIT), 2011 Second International Conference on (pp. 253-256). IEEE.

[3] Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In ACM sigmod record (Vol. 29, No. 2, pp. 93-104). ACM.

[4] E. Levent, M. Steinbach and V. Kumar, "A New Shared Nearest Neighbor Clustering Algorithm and its Applications"

[5] A.arning,R.Agrawal and P.Raghavan, A linear method for deviation detection in large database,1996.

[6] Kaur, P., & Kaur, P. AN OVERVIEW OF DATA MINING TOOLS,2015

[7] Marghny, M. H., and Ahmed I. Taloba. "Outlier Detection using Improved Genetic K-means." Available at SSRN 2545143 (2011).

[8] Shashikala, H. M., George, R., & Shujaee, K. A. (2015, April). Outlier detection in network data using the Betweenness Centrality. In SoutheastCon 2015 (pp. 1-5). IEEE.

[9] Kumar, V., Kumar, S., & Singh, A. K. Outlier Detection: A ClusteringBased Approach. International Journal of Science and Modern Engineering (IJISME), ISSN, 2319-6386.

[10]. Palla, Gergely, Imre Derényi, Illés Farkas, and Tamás Vicsek. "Uncovering the overlapping community structure of complex networks in nature and society." Nature 435, no. 7043 (2005): 814-818.

[11]. Yildiz, Hakan, and Christopher Kruegel. "Detecting social cliques for automated privacy control in online social networks." In Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on, pp. 353-359. IEEE, 2012.