Outlier Analysis in Complex Network Using DB scan Algorithm and Neural Network

Tejhaskar A[1], Vigneshwar R[2]
[1]*UG Scholar Computer Science and Engineering,*
[2]*UG Scholar, Sri Sai Ram Institute of Technology, Chennai, 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 DBSCAN clustering and neural network as novel to detect the outlier in network analysis. Especially in a social network, DBSCAN 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.**

***Index Terms*- Outlier Detection; Network Data; Adjacency Matrix; DBSCAN Clustering; Neural Network.**

I. INTRODUCTION

Data mining is a procedure of extracting useful information and ultimately understandable information from huge datasets and then using it for organization's decision making process. 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, Classification based methods, Distance base method Nearest Neighbor based methods linear method 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 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 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 DBSCAN 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 overview of basic clustering methods are discussed, Terms and definitions are given in Section III, approach of proposed algorithm and proposed methodology with algorithm is given in section IV, performance evaluation is described in Section V and conclusion is described in VI.

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

2.OVERVIEW OF BASIC CLUSTERING METHODS

There are many clustering algorithms in the literature. It is difficult to provide a crisp categorization of clustering methods because these categories may overlap so that a method may have features from several categories. Nevertheless is useful to present a relatively organized picture of clustering methods. In general, the major fundamental clustering methods can be classified into the following categories,

2.1 Partitioning methods

Given a set of n objects, a partitioning method constructs k partitions of the data, where each partition represents a cluster and k _ n. That is, it divides the data into k groups such that each group must contain at least one object. In other words, partitioning methods conduct one-level partitioning on data sets. The basic partitioning methods typically adopt exclusive cluster separation. That is, each object must belong to exactly one group. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. The general criterion of a good partitioning is that objects in the same cluster are "close" or related to each other, whereas objects in different clusters are "far apart" or very different. There are various kinds of other criteria for judging the quality of partitions. Traditional partitioning methods can be extended for subspace clustering, rather than searching the full data space. This is useful when there are many attributes and the data are sparse.

2.2 Hierarchical methods:

A hierarchical method creates a hierarchical decomposition of the given set of data objects. A hierarchical method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed. The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. It successively merges the objects or groups close to one another, until all the groups are merged into one (the topmost level of the hierarchy), or a termination condition holds. The divisive approach, also called the top-down approach, starts with all the objects in the same cluster. In each successive iteration, a cluster is split into smaller clusters, until eventually each object is in one cluster, or a termination condition holds. Hierarchical clustering methods can be distance-based or density and continuity based. Various extensions of hierarchical methods consider clustering in subspaces as well.

2.3 Density-based methods:

Most partitioning methods cluster objects based on the distance between objects. Such methods can find only spherical-shaped clusters and encounter difficulty in discovering clusters of arbitrary shapes. Other clustering methods have been developed based on the notion of density. Their general idea is to continue growing a given cluster as long as the density (number of objects or data points) in the "neighborhood" exceeds some threshold. For example, for each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points. Such a method can be used to filter out noise or outliers and discover clusters of arbitrary shape. Density-based methods can divide a set of objects into multiple exclusive clusters, or a hierarchy of clusters.

2.4 Grid-based methods:

Grid-based methods quantize the object space into a finite number of cells that form a grid structure. All the clustering operations are performed on the grid structure (i.e., on the quantized space). The main advantage of this approach is its fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space. Using grids is often an efficient approach to many spatial data mining problems, including clustering. Therefore, grid-based methods can be integrated with other clustering methods such as density-based methods and hierarchical methods.

3. 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.

Fig 1.adjacency matrix

•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.

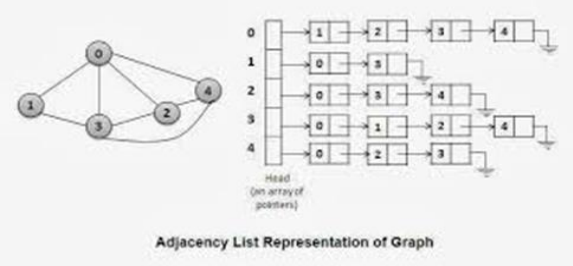

Fig 2.adjacency list

• DBSCAN

(Density-Based Spatial Clustering of Applications with Noise) finds core objects, that is, objects that have dense neighborhoods. It connects core objects and their neighborhoods to form dense regions as clusters. A user-specified parameter € > 0 is used to specify the radius of a neighborhood we consider for every object. The € -neighborhood of an object o is the space within a radius € centered at o. Due to the fixed neighborhood size parameterized by €, the density of a neighborhood can be measured simply by the number of objects in the neighborhood. To determine whether a neighborhood is dense or not, DBSCAN uses another user-specified parameter, MinPts, which specifies the density threshold of dense regions. An object is a core object if the €-neighborhood of the object contains at least MinPts objects. Core objects are the pillars of dense regions.

Given a set, D, of objects, we can identify all core objects with respect to the given parameters, € and MinPts. The clustering task is therein reduced to using core objects and their neighborhoods to form dense regions, where the dense regions are clusters.

For a core object q and an object p, we say that p is directly density-reachable from q (with respect to € and MinPts) if p is within the €-neighborhood of q. Clearly, an object p is directly density-reachable from another object q if and only if q is a core object and p is in the €-neighborhood of q. Using the directly density-reachable relation, a core object can "bring" all objects from its € neighborhood into a dense region. In DBSCAN, p is density-reachable from q (with respect to _ and MinPts in D) if there is a chain of objects p1,….. ,pn, such that p1 D q, pn D p, and piC1 is directly density-reachable from pi with respect to € and MinPts, for 1 ≤ i ≤n, pi 2 D. Note that density-reachability is not an equivalence relation because it is not symmetric. If both o1 and o2 are core objects and o1 is density-reachable from o2, then o2 is density-reachable from o1. However, if o2 is a core object but o1 is not, then o1 may be density-reachable from o2, but not vice versa. To connect core objects as well as their neighbors in a dense region, DBSCAN uses the notion of density-connectedness. Two objects p1,p2 2 D are density-connected with respect to € and MinPts if there is an object q 2 D such that both p1 and p2 are density reachable from q with respect to € and MinPts. Unlike density-reachability, density connectedness is an equivalence relation. It is easy to show that, for objects o1, o2, and o3, if o1 and o2 are density-connected, and o2 and o3 are density-connected, then so are o1 and o3.

Algorithm: DBSCAN: a density-based clustering algorithm.

Input:

D: a data set containing n objects,

€: the radius parameter, and

MinPts: the neighborhood density threshold.

Output: A set of density-based clusters.

Method:

(1) mark all objects as unvisited;
(2) do
(3) randomly select an unvisited object p;
(4) mark p as visited;
(5) If the €-neighborhood of p has at least MinPts objects
(6) create a new cluster C, and add p to C;
(7) let N be the set of objects in the €-neighborhood of p;
(8) for each point p0 in N
(9) if p0 is unvisited
(10) mark p0 as visited;
(11) if the €-neighborhood of p0 has at least MinPts points,
add those points to N;
(12) if p0 is not yet a member of any cluster, add p0 to C;
(13) end for
(14) output C;

(15) else mark p as noise;
(16) until no object is unvisited;

•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 DBSCAN 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 DBSCAN 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.

4. PROPOSED METHODOLOGY

The proposed methodology uses two methods to detect the outliers, one for eliminating the outlet node, second for outlier detection with DBSCAN 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 DBSCAN and neural network techniques and methods, which are used to detect the outlier from huge amount of data.

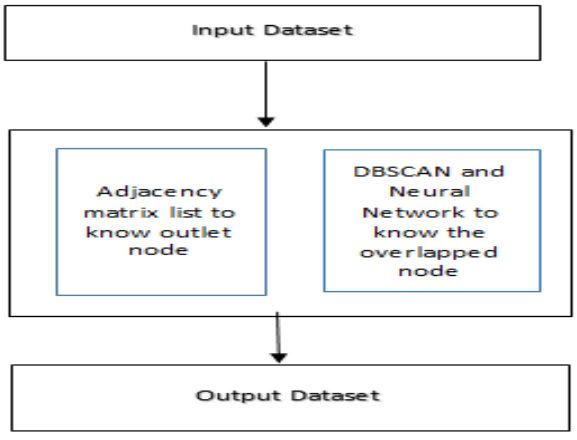

Fig 3. System architecture of proposed method

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: DBSCAN is a method of clustering. It divides the data into k numbers of clusters

Phase 3: the output of DBSACN 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 K clique 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.

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(ACi)=Ni
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 DBSCAN identify distinct cluster
n) Calculate optimal values using Fuzzy Min Max Technique

5. PERFORMANCE EVALUATION

The following graph shows the highest performance in DBSCAN when compared it with k-means algorithm and DENCLUE algorithm.

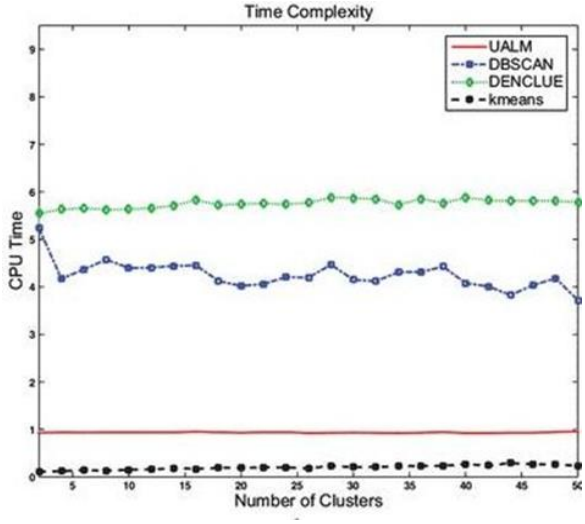

Fig 4. Performance evaluation

5.1 OTHER MEASURES

The other measures like purity, homogeneity, completeness, precision and other quality metric parameter are said to be high in DBSCAN when compared to CLARA and PAM.

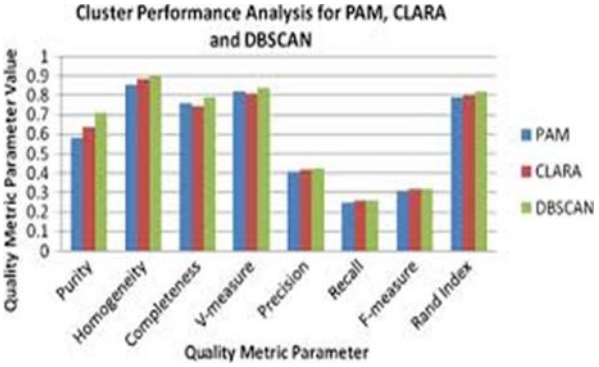

Fig 5. Quality metric parameter

6. 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.

REFERENCES

[1] Jiawei Han, Micheline Kamber, Jian Pei "Data Mining Concepts and Techniques" , ISBN: 978-0-12-381479-1.

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

[3] 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.

[4] 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.

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

[6] 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.

[7] Sanjay Chakraborty, Prof. N.K Nagwani, Lopamudra Dey "Performance Comparison of Incremental K-means and Incremental DBSCAN algorithm", International Journal of Computer Applications, ISSN, 0975-8887. Vol.27 No.11, August 2011.

[8] 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.

[9] Yaminee S. Patil, M.B. Vidhya "A technical survey on clustering analysis in data mining" International Journal of Emerging Technology and Advanced Engineering.

[10] Nidhi Suthar, Indrjeet Rajput, Vinit Kumar Gupta " A technical Survey on DBSCAN clustering algorithm" International Journal of Scientific and Engineering Research, Volume4, Issue5,May 2013.