IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.6, Issue.10, Oct. 2018 E-ISSN: 2347-2693

AN EMINENT WAY OF AN IMPROVING A DENCLUE ALGORITHM APPROACH FOR OUTLIER MINING IN LARGE DATABASE

R. Prabahari[1*] , M. Ramalingam[2]

[1,2]Dept. of Computer Science, Gobi Arts & Science College, Gobi, Tamilnadu, India

[*]*Corresponding Author: prabahari_r@rediffmail.com*

Available online at: www.ijcseonline.org

Accepted: 14/Sept/2018, Published:31/Oct/2018

Abstract- The number of methods available in data mining to detect the outlier by making the clusters of data and then detect the outlier from them. The objects that are similar to each other are organized in group it's called cluster and the objects that do not comply with the model or general behavior of the data these data objects called outliers. Outliers detect by clustering. Density based clustering algorithm (DENCLUE) is one of the primary methods for clustering in data mining which groups neighboring objects into clusters based on local density conditions rather than proximity between objects. Data points are assigned to a cluster by hill climbing, points going to the same local maximum are put into the same cluster. The traditional density estimation only considers the location of the point, not variable of interest. Depending on the convergence criteria, the method needs less iteration as fixed step size methods and improving cluster quality and also finding an outlier correctly.

***Keywords*:** Clustering, Data Mining, Density Based Clustering Algorithm, DBSCAN, OPTICS, Outlier Mining,

I. INTRODUCTION

Data mining is the process of extraction of hidden patterns or characteristics from such large datasets and transforms it in an understandable manner. There are many tasks involved in data mining. One of the tasks is clustering where a set of objects is divided into several clusters where the intra-cluster similarity in maximized and the inter-cluster similarity in minimized.

Big data may be defined as a term for data sets that are so large or complex that traditional data processing applications prove inadequate. Big data exhibits different characteristics like volume, variety, velocity and veracity due to which it is very difficult to analyze data and obtain information with traditional data mining techniques [1].

Figure 1. Clustering Stages

Clustering algorithms may be broadly classified into partition-based algorithms, density-based algorithms, hierarchical-based algorithms and grid-based algorithms. The DBSCAN, OPTICS and DENCLUE are some of the most commonly used density-based clustering algorithms[2]. Density based algorithms find the cluster according to the regions which grow with high density. There are two approaches that may be used in density-based methods. The first approach, called the density-based connectivity clustering. The algorithms that represent this behavior include DBSCAN and OPTICS. The second approach pins density to a point in the attribute space and is called Density Functions. This behavior include algorithm DENCLUE.

Outliers are objects in the data, which are significantly different from the rest of the data. Outlier detection is a process of finding such anomalies in the data. Outlier detection is a primary step in many data-mining applications.

Section I contains the introduction of clustering and outlier data. Section II contains the related works of clustering and outlier mining. Section III deals with the problem of existing implementation of Dbscan and Optics algorithm. Section IV specifies the objectives of the proposed algorithm.. Section V explain the methodology of the proposed algorithm. Section VI describes results and discussion. Section VII specify concludes research work with future directions.

II. RELATED WORKS

Data objects or elements that are entirely different from others or inconsistent in comparison to other data elements are referred to as Outliers. Outlier data do not comply with the general behavior of the database and exhibit deviant and aberrant behavior. It could also be viewed as the process of clustering, but with the difference that clusters look out for the objects or records that have the least similarity and different behavior compared to the rest of the data. Mining for outliers is an important data mining research process with numerous applications including credit card fraud

detection [3], identifying computer network intrusions [4], detecting employers with poor injury histories [5], discovery of criminal activities in e-commerce, weather prediction, marketing and customer segmentation. Two key phases of outlier mining are: identifying the inconsistent data in the large input database and the extraction of the expected number of outliers or deviant data points. The commonly used approaches are statistical based approaches [6], distance based approaches, cluster based approaches and density based approaches [7]. The Statistical model starts out with a distribution or probability model for the given data set and then looks out for deviation from the considered model. Distance based technique carries forward the concept used in clustering, with the modified objective of grouping or looking out for data points that lie in far distances. In unsupervised learning, outliers that are considered as noise [8] because they can severely affect the results of clustering. It is removed from the analysis. In clustering based methods, the normal data objects belong to large and dense clusters, whereas outliers belong to small or sparse clusters, or do not belong to any clusters. Density based outlier [9] detection is closely related to distance based outlier detection since density is usually defined in terms of distance and this paper deals with density based approach to find cluster[10]. Alexander Hinneburg et al [11], proposed a new algorithm for clustering in large multimedia databases i.e. called DENCLUE that can handle noise. In this approach, they are able to find nonspherical shaped clusters using local density function. They evaluated performance of DBSCAN with DENCLUE which shows that DENCLUE is more superior to DBSCAN. B.G. Obula Reddy et al., [12] proposed a comparative analysis of various clustering techniques that enables us to choose best clustering algorithm by explaining each of them with characteristics, examples, positive and negative aspects. Mariam Rehman et al.,[13] provided comparison between DBSCAN and RDBC algorithm by implementing them using iris data set that concluded that RDBC is more efficient algorithm than DBSCAN as it can handle outliers more effectively. Henrik Bäcklund et.al., [14], has first provided description about DBSCAN algorithm with all its relevant terminologies. Anoop Kumar Jain et al.,[15] presented a survey of recent clustering techniques for data mining research that includes centroid based, connectivity based, density based and distributive based clustering and discussed about k-means,rapid clustering method and hierarchical agglomerative clustering. Rui Xu et.al.,[16] explained procedure of cluster analysis using few steps with a feedback loop. They carried out complexity comparison among various clustering algorithms.

III. PROBLEM STATEMENT

This algorithm is concerned with analyzing and finding the solutions of the problem of density clustering in order to find outliers. The density based clustering is influenced by the density divergence problem that affects the accuracy of clustering and cannot choose its parameter according to the distribution of the data set [17]. The traditional density estimation is also considered as the location of the point, not variable of interest and hill climbing may create unnecessary small steps in the beginning and never converges exactly to the maximum. The solution produced by the existing algorithms like DBSCAN and OPTICS are not effective on the above issues [18].

IV OBJECTIVES

For Knowledge Discovery in Database (KDD) applications, finding the outliers, i.e. the rare events, is more interesting and useful than finding the common cases. The main objectives of this algorithm have been

- To improve cluster quality and finding correct outlier.
- To produce more accuracy in hill climbing process to find density attractors

V. METHODOLOGY

The overall problem can be formulated as follows:
Assume that datasets have the form (<Location>, <variable_of_interest>). More formally, a dataset *O* is a set of data objects, where n is the number of objects in *O* belonging to a feature space F.

$$O = \{o_1, o_2, o_3 \ldots o_n\} \varepsilon F \quad (5.1)$$

Assuming that objects $o \varepsilon O$ have the form $((x, y), z)$ where (x, y) is the location of object *o*, and *z* denoted as $z(o)$ is the value of the variable of interest of object *o*. The variable of interest can be continuous or categorical. Besides, the distance between two objects in *O*, $o_1 = ((x_1, y_1), z_1)$ and $o_2 = ((x_2, y_2), z_2)$ *is* measured as $d((x_1, y_1), x_2, y_2))$ where *d* denotes a Euclidian distance[19].

5.1 Influence and Density Functions

In general, density estimation techniques employ influence functions that measure the influence of a point $o \varepsilon O$ with respect to another point $v \varepsilon F$, a point *o*'s influence on another point *v*'s density decreases as the distance between *o* and *v* increases. In particular, the influence of object $o \varepsilon O$ on a point $v \varepsilon F$ is defined as:

$$f_{influence}(v, o) = z(o) * e^{*(-d(v, o)^2)} / (2\sigma^2) \quad (5.2)$$

If for every $o \varepsilon O$, $z(o) = 1$ holds, the above influence function become a Gaussian kernel function.The parameter σ determines how quickly the influence of *o* on *v* decreases as the distance between *o* and *v* increases. The overall influence of all data objects $o \varepsilon O$ on a point $v \varepsilon F$ is measured by the density function $\psi^o(v)$, which is defined as follows:

n

$$\Psi^o(v)=\sum_{i=1} f_{influence}(v_i, o_i) \quad (5.3)$$

5.2 Local maximum procedure

The improved algorithm operates on the top of the influence and density functions that were introduced in equations 5.1, 5.2 and 5.3. Its clustering process is a hill-climbing process that computes density attractors. During the density attractor calculation process, data objects are associated with density attractors forming clusters. A point a, is called a density attractor of a dataset *O* if and only if it is a local maximum or minimum of the density function ψ^o and $|\psi^o(a)| > \xi$, where ξ is a density threshold parameter. For any continuous and differentiable influence function, the density attractors can be calculated by using a hill climbing procedure [20]. The proposed method does not calculate the density attractor for every data object. During the density attractor calculation, the data objects closed to the current point are examined when moving towards the supervised density attractor. If their density values have the same sign as the current density value at the calculating point, those data objects will be associated with the same final density attractors after it has been computed. In each iteration step, locate the data objects close to o_i^{j+1}, i.e. data objects whose distance to o_i^{j+1} is less than $\sigma/2$. Each data object close to o_i^{j+1} is processed as follows: If the data object does not belong to any cluster yet, the object is marked with the same cluster ID as those attracted by the current attractor computation. If the data object already belongs to a cluster c_i, all the data points in the cluster c_i *are* marked with the same cluster ID as those objects that have been collected by the current attractor computation. The hill climbing procedure stops returning a density attractor *a*, if $|\psi^o(a)| > \xi$, a is considered a density attractor, and a cluster is formed.

An improved hill climbing procedure performs the following steps. First, the influence and density function is used. The influence of each data point can be modeled formally using a mathematical function which is called an influence function. Influence function describes the impact of data point within its neighborhood and then the density function calculated. Secondly, clusters can be determined by identifying density attractors where density attractors are local maximum of the overall density function which is the sum of influences of all data points. Finally, the outlier can be detected if any point does not belong to cluster. The improved algorithm is shown in the Figure 2.

Influence & Density function

Step1: find the distance of dataset1(x) and dataset2(y).
Step2: find I(x,y) = z(o) * exp { - [distance(x,y)**2] / [2*(sigma**2)] }(σ, the std. dev.).
Step3: Repeat the step4 until end of the data set to find sum of influence of each another dataset.
Step4: Density = Density + Influence (entity, sigma).

Density-attractor

Step1: Take arbitrary any object x_i.
Step2: Repeat step3 until find local density attractor **(**x_n < density threshold).
Step3: Calculate gradient to find neighbor point.
Step4: Move to next point x_{i+1}.

Clustering and outlier

Step1: Data Set O ={ o_1, o_2, o_3, o_4, ... o_n}.
Step2: Find (Highly) Populated Cells (use a threshold=ξc)
Step3: Identify populated cells (+nonempty cells).
Step4: For any uncluster data objects, find density attractor points, C*, using hill climbing iteratively:
(i) Randomly pick a point, p_i.
(ii) Compute local density (r=4σ).
(iii)Pick another point, p_{i+1}, close to p_i, compute local density at p_{i+1}.
(iv)If LocDen(p_i) < LocDen(p_i+1), Climb.
(v) Put all points within distance σ/2 of path, p_i, p_{i+1}, …C* into a density attractor cluster called C*.
Step5: Connect the density attractor clusters, using a threshold, ξ, on the local densities of the attractors.
Step6: If any object does not belong to C*, it is called outlier object.

Figure 2: Outlier Mining Algorithm

VI. RESULTS AND DISCUSSIONS

In order to show the different density based clustering algorithms and DENCLUE algorithm is used in synthetic and benchmark image dataset. Four bench mark datasets (Iris, Liver Disorder, Breast Tissue, Spectf Heart) are used and the datasets taken from the machine learning repository at UCI. The Iris dataset is multivariate and has three classes consisting of 50 objects. The liver disorder dataset has 345 instances and 7 features. The Breast Tissue dataset has 106 object and 10 features and Spectf Heart dataset has 267 ojects and 22 features.

Experiment is done on three parameters in order to compare DBSCAN, OPTICS, DENCLUE. First parameter is shape of clusters. All the three algorithms support arbitrary shape of clusters. Second parameter is Handling of Noise as noise increases DENCLUE performs very well and OPTICS also perform well but in case of DBSCAN, it does not perform so well. Third one is Cluster quality that is defined in terms of F score follows. So, DENCLUE is superior to DBSCAN and OPTICS.

Table 1: Comparison of Density based Clustering Algorithms

Algorithms	Clusters Shape	Noise Handle	Cluster Quality
DBSCAN	Arbitrary	Not good	91.3%
OPTICS	Arbitrary	Good	94.3%
DENCLUE	Arbitrary	Very Good	97.08%

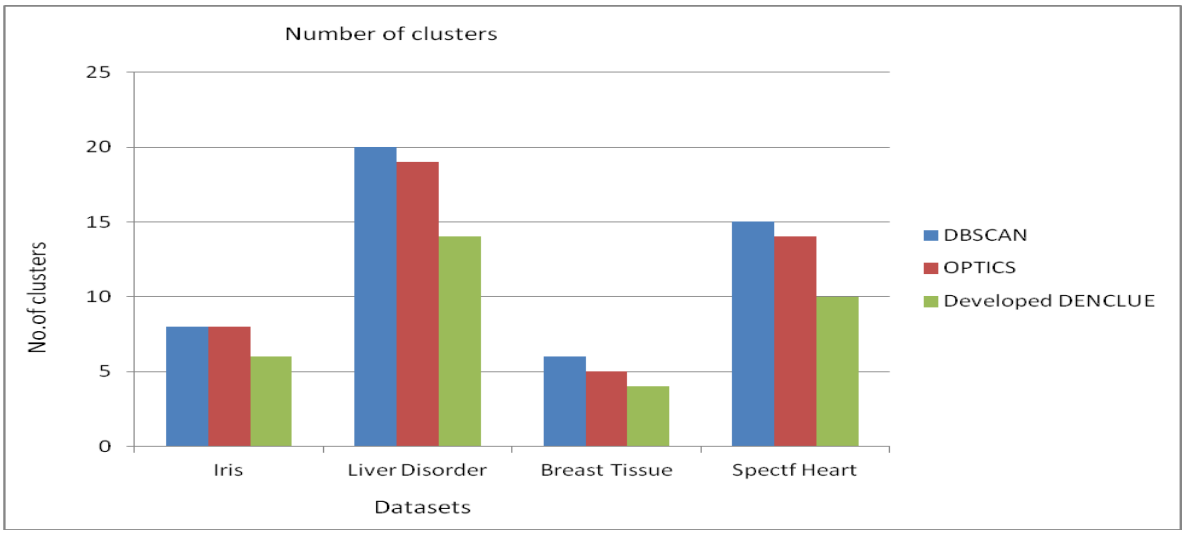

Figure 3. Performance comparisons between existing and proposed DENCLUE on number clusters formed in different datasets

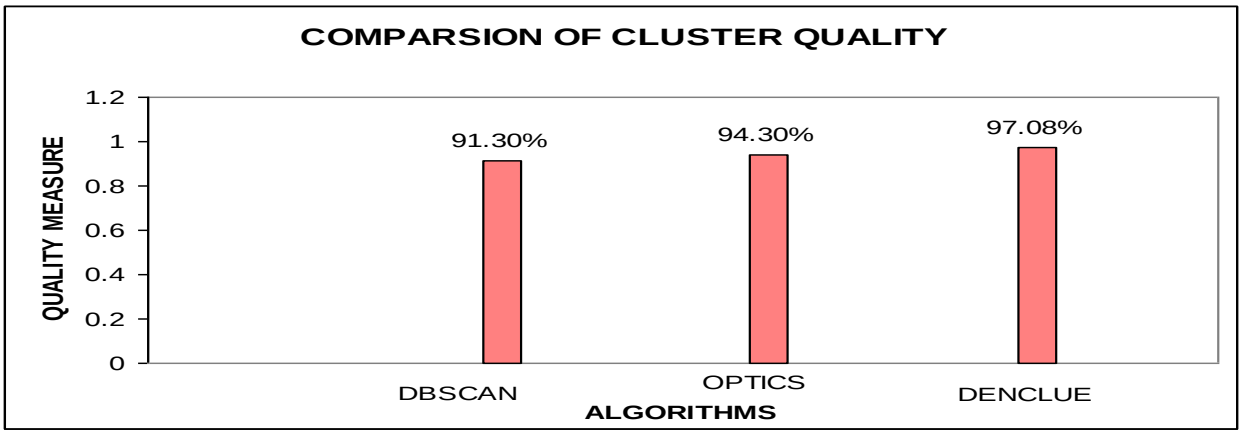

Figure 4. Comparison of algorithms based on quality

In terms of cluster quality DENCLUE leads while OPTICS and DBSCAN is lacking behind. From the comparison, the numbers of clusters are formed in developed DENCLUE algorithm is less than to traditional methods. Due to the reduction in number of clusters the outlier can be found correctly.

VII. CONCLUSION

This improved density based clustering approach that extends the traditional density estimation techniques by considering a variable of interest that is associated with a spatial object. Density is measured as the product of an influence function with the variable of interest. The algorithm uses local maximum method to calculate the maximum (density attractors) of a density function and clusters are formed by associating data objects with density attractors during the local maximum procedure. Analyses the outlier object and get some knowledge from the objects because the outlier can be useful for many applications and also further improved to reduce the run time. Compared with other algorithms, the improved algorithm produces significantly better results.

REFERENCES

[1]. Harsh Shah, Karan Napanda and Lynette D'mello, "Density Based Clustering algorithm", International Journal of Computer Engineering, vol. 3,Issue. 11, pp.54-57, Nov 2015, E-ISSN: 2347-2693

[2]. Anoop Kumar Jain, Prof.Satyam Maheswari,"Survey of Recent Clustering Techniques in Data Mining", International Journal of Computer Science and Management Research ,Vol. 1, Issue 1, Aug 2012.

[3]. R.J Bolton, D.J.Hand, "Statistical fraud detection": A review with discussion, Statistical Science, 17(3): pp. 235-255, 2002.

[4]. E. Eskin , A.Arnold , M.Prerau , L.Portnoy , S.Stolfo ," A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data", In Data Mining for Security Applications, 2002.

[5]. E. Knorr, R.Ng., V. Tucakov, ," Distance-based outliers: Algorithms and applications", The International Journal on Very Large Data Bases Journal 8, pp. 237–253, 2000.

[6]. J.Laurikkala, M. Juhola and E. Kentala, "Informal Identification of Outliers in Medical Data", in Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology IDAMAP-2000 Berlin,Organized as a workshop of the 14th European Conference on Artificial Intelligence ECAI-2000.

[7]. M.Breunig, H.Kriegel,R. Ng , J.Sander, "LOF: Identifying density based local outliers", In: Proc. SIGMOD Conf, pp. 93–104, 2000.

[8]. C.Aggarwal,P.S. Yu , "Outlier Detection for High Dimensional Data". In: Proceedings of the ACM SIGMOD Conference 2001.

[9]. Anant Ram, Sunita Jalal, S. Anand . Jalal, Manoj kumar, "A density Based Algorithm for Discovery Density Varied cluster in Large spatial Databases", International Journal of Computer Application Vol. 3,No.6, June 2010.

[10]. R.Prabahari ,V.Thiagarasu , "A Comparative Analysis of Density Based Clustering Techniques for Outlier Mining", International Journal Of Engineering Sciences & Research Technology, ISSN 2277-9655, pp 132-136 November, 2014.

[11]. Alexander Hinneburg, Daniel A.Keim (1998),"An Efficient Approach to Clustering in Large Multimedia Databases with Noise [Online] Available: http://www.aaai.org

[12]. B.G Obula Reddy, Dr. Maligela Ussenaiah,"Literature Survey On Clustering Techniques", IOSR Journal of Computer Engineering, Vol. 3, Issue 1, July 2012.

[13]. Mariam Rehman, Syed Atif Mehdi,"Comparison of Density Based Clustering Algorithms", research work, Lahore College for Women University, Lahore, Pakistan.

[14]. Henrik Bäcklund, Anders Hedblom, Niklas Neijman, "DBSCAN A Density-Based Spatial Clustering of Application with Noise", 2011.

[15]. Anoop Kumar Jain, Prof.Satyam Maheswari,"Survey of Recent Clustering Techniques in Data Mining", International Journal of Computer Science and Management Research ,Vol. 1, Issue 1, Aug 2012.

[16]. Rui Xu, Donald Wunsch,"Survey of Clustering Algorithms", IEEE Transactions On Neural Netwoks, Vol. 16, No. 3, May 2005

[17]. R.Prabahari ,V. Thiagarasu **,** "Density Based Clustering Using Gaussian Estimation Technique" , International Journal on Recent and Innovation Trends in Computer Science and Communication(IJRITCC), ISSN 2321-8169, pp 4078-4081 December, 2014.

[18]. Jianhao Tan and Jing Zhang "An Improved Clustering Algorithm Based on Density Distribution Function" Computer and Information Science Vol. 3, No. 3; August 2010.

[19]. He Zengyou, Xu Xiaofei , Deng Shengchun, Squeezer, "An efficient algorithm for clustering categorical data", Journal of Computer Science and Technology, pp. 611-624, May 2002.

[20]. Shweta Verma, Vivek Badhe "Survey on Big Data and Mining Algorithm" International Journal of Scientific Research in Science, Engineering and Technology, Vol. 2 ,May 2016 Online ISSN : 2394-4099.

Authors Profile

Dr.R.Prabahari, is an Assistant Professor in the Department of Computer Science, Gobi Arts & Science College, Gobichettipalayam. She has 11 years of experience in teaching. She has published more than 10 papers in international journals and conferences. Her Research area Data Mining and Warehousing and focuses on outlier mining for discovering abnormal and irregular patterns of images efficiently, by improving cluster quality. She is a Life Member in Indian Science Congress.

Dr. M.Ramalingam is an Assistant Professor in the Department of Computer Science, Gobi Arts & Science College, Gobichettipalayam. He has 4 years of relevant industrial experience and decade of rich experience in research and teaching. He has published more than 25 papers in international journals and conferences. His Research focuses on Cluster based Mobile Ad hoc Network and MANET security. He is a Life Member in Indian Science Congress, Computer Science Teachers Association (CSTA), Society of Digital Information and Wireless Communications (SDIWC), Internet Society (ISOC) and Member in The Global Community of Information Professionals.

