RESEARCH ARTICLE OPEN ACCESS

Comparative Analysis of Outlier Detection in Clustering Algorithm Using Optimization Techniques

[1] B.Angelin , [2] Dr.A.Geetha ,
[1]Research Scholar (Ph.D), [2]Assistant Professor & Head in Computer Science,
[1, 2,] Pg&Research Department of Computer Science,
[1, 2] Chikkanna Government Arts College, Tirupur.
Mail : angelindia93@gmail.com , gee_sam@yahoo.com

--************************----------------------------------

Abstract:

Clustering is the task of assigning a set of data objects into groups called clusters so that the objects in the same cluster are more similar in some sense to each other than to those in other cluster. Data items whose values are different from rest of the data or whose values fall outside the described range are called outliers. Outlier detection is an important issue in data mining, where it is used to identify and eliminate anomalous data objects from given data set. This paper provides a brief on clustering techniques and outlier detection techniques and optimization techniques. Particularly the ROC Curve, K means and k median clustering algorithm, dragonfly optimization algorithm, particle swarm optimization algorithm for outlier detection is discussed.

Keywords: outlier data mining, multi objective PSO, ROC,hybrid dragon fly.

--************************----------------------------------

I. INTRODUCTION

Clustering is an important technique in data analysis and data mining applications. Clustering is the partition of data into groups in a way that objects in the same group are more similar to each other and different from objects of other groups. These groups are called clusters.[2][3] Data mining is an activity which is carried out in different steps. These are anomaly detection, association, classification and clustering. Data mining is the searching and training of large data sets, in order to find out significant patterns and rules. Data mining is one of the best ways to illustrate the difference among data and information. It converts data into information.[12][14] A good clustering method will produce high quality clusters with high intra-clusters parallel and low inter cluster parallel. In well-separated clusters points are nearer to every other point in the cluster than to any point not in the cluster. Center based: It resides then the other clusters then it is named as center based cluster. Contiguous cluster:[4][5] Nearer to one or more other points in the cluster than to any point not in the cluster. Density based cluster: Low density states, from other regions of high density. This type of cluster used only when the clusters are irregular or intertwined and when sound outliers are present. Shared property-share some common property or represents a particular notion,There are two learning approaches used, supervised learning and unsupervised learning.

II. LITERATURE SURVEY

B.Angelin [2021] develop a system that purpose Dragon fly K-means algorithm is to discover the optimum information with the larger records and smaller attributes with high classification rate. The high classification or detection rate is very important for good statement and intrusion detection. This objective is achieved by mistreatment the mythical monster property negative quantitative relation of the classifier as its objective or fitness operate for dragon fly k-means algorithm. In this, the planned methodology is tested on the clinical datasets which needs high classification rate for discover the wellness at associate degree early stage.[6][8]

Rajendra Pamula (2011) is plan a proposed framework a clustering based strategy to catch outliers . We apply K-implies clustering algorithm to isolate the data index into bunches.[1] The focuses which are lying close to the centroid of the bunch are not plausible contender for exception and we can prune out such focuses from each group. Next we ascertain a distance based exception score for outstanding focuses. The calculations expected to figure the outlier score diminishes extensively because of the pruning of certain focuses. In light of the exception score we pronounce the top n focuses with the most noteworthy score as outliers. The exploratory outcomes utilizing genuine informational index exhibit that despite the fact that the quantity of calculations is less, the proposed technique performs better compared to the current strategy.[11][13][10]

B. Angelin (2020) is build up a proposed framework The valuation of anomaly recognition algorithm be a steady debate in information mining examine. Little is known in regards to strength and shortcoming of disparate customary outlier recognition technique s, in this paper We play out a broad exploratory investigate the exhibition of an agent set of enhancement based exception identification across a tremendous sort of datasets arranged consequently dependent on the general execution of the exception location strategies, we likewise propose extra fitting for the assessment of exception discovery results. We utilized dragonfly and k means calculations and advancement strategy used to identify the outlier dependent on ROC Curve technique.[7][9]

Outlier detection using K means and K median algorithm: The mean in the k-median grouping, and the median is determined for each measurement in the information vector. The centroid of a group can be characterized by unique routes in the k-median cluster, the median has utilized the mean. The information preserve is unreservedly utilized for nonprofit influencing instruction or research purpose. Center for Machine Learning and Intelligent Systems with a predicted attribute as iris plant. The proposed method also located the outliers with a small number of points with different colors. The significance of all the above experiments can't be overlooked. Overall, all the optimizers have successfully brought about an appreciable improvement in reaching better centroids after a few iterations. As the sizes of data sets continue to grow rapidly these days with easy availability of relevant data for analysis.

Iris Dataset	Number of Outliers			
	Existing Enhanced K-Means Clustering Algorithm		Proposed Enhanced K-median Outlier Clustering Algorithm	
Type	Petal	Sepal	Petal	Sepal
Values	2.0067	2.0067	1.9533	2.0200

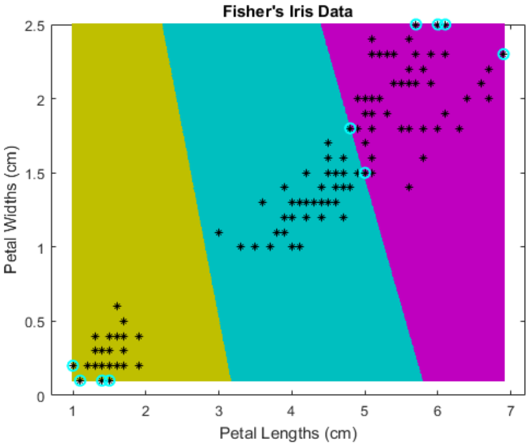
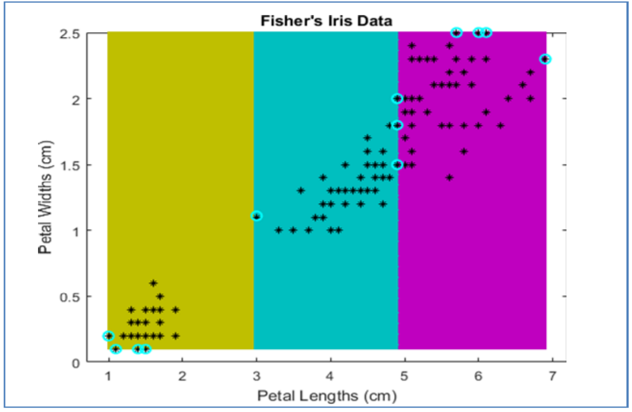


Figure 1: Petal Outlier Detection by K means
Figure 2: Petal Outlier Detection by K median



The Method is based on the K-means clustering based outlier detection and K-median based outlier detection. In K-median outlier detection, the following steps are followed

1. The dataset is divided into clusters using divide and conquer algorithm.

2. Then, it sort the elements based on the distance between them to group the elements in the cluster.

3. The data omitted from the grouping of the clustered elements is considered as the outlier element for the K-median clustering.

4. The process is repeated for all clusters.

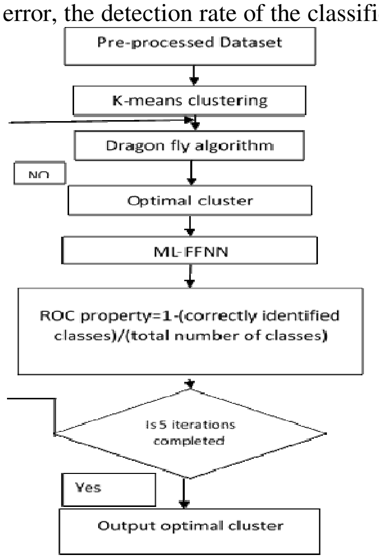
The drawback in this approach it determine only the outlier which perform repeated distance calculation to detect the outliers. To overcome this drawback, a dragon fly based k-means clustering is proposed for the outlier detection and feed forward neural network is used for the classification. The brief explanation of the proposed method is given in the section ,Objectives of the proposed Method: In this, the objectives of the Dragon fly K-means clustering is to overcome the drawbacks of the existing method.

The other objectives of the proposed method are as follows:

- To determine the optimal cluster.
- A repeated median based outlier detection is proposed.
- Improve the classification rate of the application.

Roc curve based outlier detection using dragonfly optimization:

The main objective of the proposed Dragon fly K-means clustering is to detect the optimal data with the larger records and smaller attributes with high classification rate. The high classification or detection rate is important for perfect forecasting and intrusion detection. This objective is achieved by using the roc property negative ratio of the classifier as its objective or fitness function for dragon fly k-means clustering. In this, the proposed method is tested on the clinical datasets which requires high classification rate for detect the disease at an early stage. The optimal data from the dragon fly based K-means clustering is used for the classification and also to evaluate the outlier detection performance. The classification process is performed using the multi-layer feed forward neural network [23]. The multi-layer feed forward neural network consists of three layers namely Input layer, hidden layer and output layer. The Multi-layer feed forward neural network uses the back propagation algorithm which is useful for training by reducing the mean square error between the outputs and targets which is propagated backwards from the output layer to the input layer. By reducing the mean square error, the detection rate of the classification is increased.



Parameter	Elliptical seizure	IRIS flower
No of records	11500	150
No of attributes	178	4
Actual classes	5	3
Converted classes	2	3
Existing techniques applied	No	No

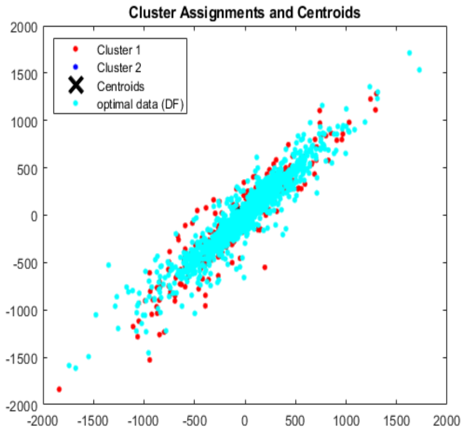


Figure :3 Elliptical seizure DF_K-Means clustering

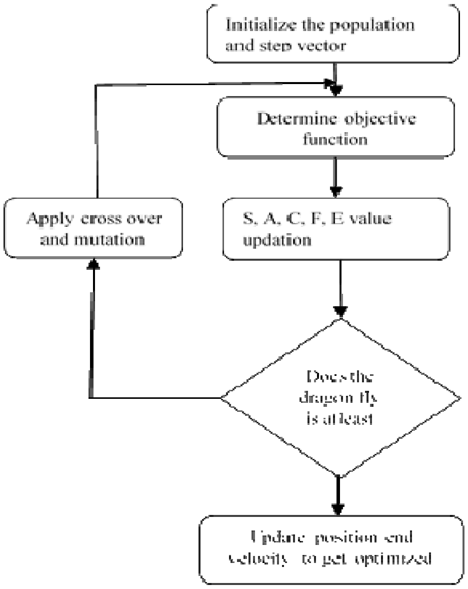
Comparison of outlier detection approaches:

Dataset	Proposed
Elliptical seizure	0.9649
Iris dataset	0.9759

These technique able to process the larger dataset like 11,500 records with 178 attributes of Elliptical seizure dataset and 4 attributes of the Iris flower dataset. The dragon fly based clustering able to classify the elliptical seizure dataset with 0.96 detection rate and 0.9759 for the Iris data set. Based on the performance evaluation it is observed that the proposed outlier detection is able to process all types of datasets irrespective of sizes. In future the proposed method can be extended by varying the optimization technique to reduce the computational time for the larger datasets and to process larger records.

Roc curve based outlier detection using PSO optimization:

One of the current improvement approaches is Particle swarm optimization. The current methodology fails to manage the greater records and more humble properties. To overcome this problem a dragon fly based K-means clustering and multi-layer feed forward neural network is used in this existing method. This objective is refined with the help of the ROC curve.


Initialize the population and step vector
Determine objective function
S, A, C, F, E value updation
Apply cross over and mutation
Does the dragon fly is at least
Update position and velocity to get optimized

Flow chart for Hybrid Dragonfly

Parameter	Elliptical seizure	IRIS flower
Number of records	11500	150
Number of attributes	178	4
Actual number of classes	5	3
Converted number of classes	2	3
Existing techniques applied	NO	No

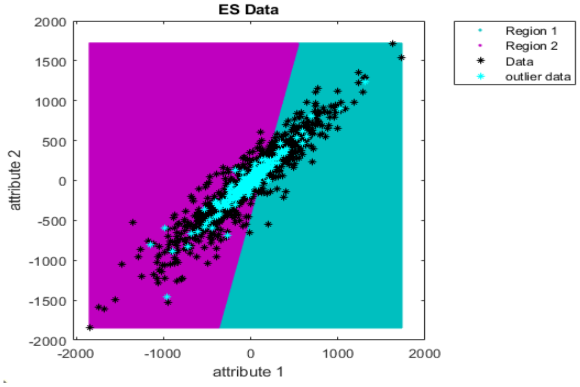


Figure: 4 Outlier detection in Elliptical seizure

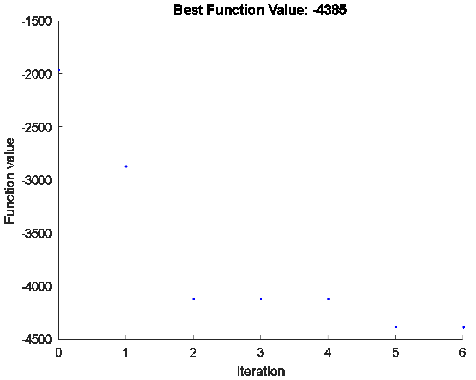


Figure:5 PSO convergence curve for elliptical seizure

CONCLUSSION

Outlier detection is an important process in data mining process to remove the data which are produced due to the manual error or disturbance. It helps in various applications like weather forecasting, performance analysis and intrusion detection. Several techniques were proposed to perform the outlier process. But they face a problem in determine the perfect data with high detection rate and low computation time. One of the best approach for outlier detection is K-means and K-median clustering. But, it failed for the larger datasets due to the repeated distance calculations. In this, it is overcome by the proposed dragon fly k-means clustering along with the proposed distance calculation to perform the outlier detection. The proposed technique able to process the larger dataset like 11,500 records with 178 attributes of Elliptical seizure dataset and 4 attributes of the Iris flower dataset. The dragon fly based clustering able to classify the elliptical seizure dataset with 0.96 detection rate and 0.9759 for the Iris data set. Based on the performance evaluation it is observed that the proposed outlier detection is able to process all types of datasets irrespective of sizes. This technique is take to more time for executing, The proposed technique is particle swarm optimization is used to detect the outlier in same dataset, this result is effective in compared with previous techniques. In future proposed method can be extended by varying the optimization technique to reduce the computational time for the larger datasets and to process larger records.

References:

1. Tan, P. N., Steinbach, M., & Kumar, V. (2016). Introduction to data mining. Pearson Education India.
2. Bansal, R., Gaur, N., & Singh, S. N. (2016, January). Outlier detection: applications and techniques in data mining. In 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence) (pp. 373-377). IEEE.
3. Zhang, Y., Hamm, N. A., Meratnia, N., Stein, A., Van De Voort, M., & Havinga, P. J. (2012).

Statistics-based outlier detection for wireless sensor networks. International Journal of Geographical Information Science, 26(8), 1373-1392.

4. Wahid, A., & Rao, A. C. S. (2019). A distance-based outlier detection using particle swarm optimization technique. In Information and Communication Technology for Competitive Strategies (pp. 633-643). Springer, Singapore.
5. Gupta, R., & Pandey, K. (2016). Density based outlier detection technique. In Information Systems Design and Intelligent Applications (pp. 51-58). Springer, New Delhi.
6. Angelin, B. "A Roc Curve Based K-Means Clustering for Outlier Detection Using Dragon Fly Optimization." *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12.9 (2021): 467-476.
7. Angelin, B., and A. Geetha.(2020) "DRAGONFLY K-MEANS FOR OUTLIER DETECTION USING CURVE METHOD." International Research Journal of Modernization in Engineering Technology (IRJMET) 2(7), ISSN:2582-5208, PP.1070-1076.
8. Angelin B., and A.Geetha, “Outlier Detection Using Clustering Techniques-K-means and K-median” Proceedings of the International Conference on Intelligent Computing Control Systems(ICICCS 2020) IEEE Xplore PP. 373-378.
9. Angelin B, and A.Geetha, “Review on Outlier detection using clustering algorithm” Science,Technology and Development , ISSN:0950-0707, 9(1) 2020, PP.374-380.
10. Duraj, A., & Chomątek, Ł. (2017). Outlier detection using the multiobjective genetic algorithm. Journal of Applied Computer Science, 25(2), 29-42.
11. 20. Duraj, A., Niewiadomski, A., & Szczepaniak, P. S. (2018). Outlier detection using linguistically quantified statements. International Journal of Intelligent Systems, 33(9), 1858-1868.
12. 21. Jayanthi, P., Kb, N. D., & Kc, K. (2016). An enhanced cuckoo search approach for outlier detection with imperfect data labels. Int J AdvEngg Tech/Vol. VII/Issue II/April-June, 667, 673.
13. 22. Mirjalili, S. (2016). Dragonfly algorithm: a new meta-heuristic optimization technique for solving singleobjective, discrete, and multi-objective problems. Neural Computing and Applications, 27(4), 1053-1073.
14. 23. Svozil, D., Kvasnicka, V., & Pospichal, J. (1997). Introduction to multi-layer feed-forward neural networks. Chemometrics and intelligent laboratory systems, 39(1), 43-62