A Roc Curve Based K-Means Clustering for Outlier Detection Using Dragon Fly Optimization

B.Angelin[a], and Dr.A.Geetha[b]

[a]
Research Scholar (Ph.D), Pg&Research Department of Computer Science,
Chikkanna Government Arts College, Tirupur
[b]Assistant Professor & Head in Computer Science, Pg&Research Department of Computer Science, Chikkanna Government Arts College, Tirupur.

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: Outlier detection is an essential step in the data mining process. Its main purpose to remove the incompatible data from the original data. This process helps in the removal of data which are necessary for carrying out to speed up the applications like classification, data perturbation and compression. It plays an important role in the weather forecasting, performance analysis of sports person and network intrusion detection systems. The outlier for the single variable can be easily observed but for the n-variable it become a tedious process. To enhance the performance of outlier detection in n-variable or attributes several methods were proposed. Some of the existing methods are statistical approaches, proximity-based measures, classification approaches, and index-based approaches and optimization based approaches. The first four approaches were not able to classify the data when there is an imperfection in the labels. But, the optimization based approach is able to overcome this problem even there is an imperfect labelling. One of the existing optimization approach is K-means and K-median based approach. The existing method failed to process the larger records and smaller attributes. To overcome this problem a dragon fly based K-means clustering and multi-layer feed forward neural network is proposed. This objective is achieved with the help of ROC curve (negative ratio) as objective function. The performance is evaluated using detection rate. The proposed method is tested on datasets like Arrhythmia, Diabetics and Epileptic seizure which has larger attributes and larger records. The proposed method outperforms the existing approach with high detection rate of 0.95 for arrhythmia dataset and 0.96 for Epileptic dataset.

Keywords: outlier data mining, k-means optimization, ROC dragon fly.

1. Introduction

Data Mining is a vast field to explore and exploit the larger datasets to determine a meaningful pattern or rules. It is differ from the normal prediction because it find out the future outcomes based on the mining process. The role of data mining is important in the following field like card fraud detection, bio medical field, Intruder detection in networks and Qualitative data assessment. The advantages of data mining are automated decision making, predicting the future results and cost reduction. But it also faces problems like big data sets, over fitting models and privacy and security [1].

The big data problem can be overcome by using the cloud and artificial intelligence techniques. The privacy and security problem is overcome by the advanced encryption techniques. But the major drawback is the over-fitting models. Because in over-fitting models the larger dataset training results in mere prediction and smaller datasets training results in false prediction. This problem can be overcome with the help of Outlier detection.

Outlier detection is a process of detecting the irrelevant information that differ from the remaining dataset. Outlier is also defined in two ways which is mentioned in [2] as follows. First definition of outlier is defined as the appearance of data which is differed from the remaining set of data. Second definition of Outlier is an observation which appears to be an inconsistent to the remaining set of data.

Generally, the outliers are classified into three types namely point outlier, Context outlier and collective outlier. Point outlier means a data point which differ from the remaining set of data. Context outlier is one in which the behaviour or the attributes of the data is differed. Collective outlier is one in which a group of data is differed from the other groups of data.

The detection of outlier is classified into two types as classic outlier and spatial outlier. The classic outlier has four categories namely statistical based approach, deviation based approach, distance based approach and density based approach. The spatial outlier detection has two categories namely space based and Graph based approach [2].

The statistical outlier detection is used in the Wireless sensor networks for the node information and intruder detection [3]. The statistical approach is applied on determining the correlation property on the temporal and spatial properties of the WSN data. [4] Employed the PSO based distance based outlier detection for datasets like Yeast, Ionosphere etc., In that, the particle swarm optimization is used to detect the anomaly based on the distance between the data points in the Knn grouping.

The density and statistical measure based outlier detection is performed on the breast cancer dataset and Iris dataset [5]. There are different techniques used in density based outlier detection. Each technique has individual properties and it is used for different applications. In [5], the standard deviation property is used for determine the outlier in medical and natural dataset.

A Fractal based outlier detection is proposed for determining the outliers in radio frequency data streams [6]. In this, the outlier is detected by dividing the data streams into multiple small fractal based on time. By sorting the fractal and repeating its process over different sliding windows is used for detect the outliers. Another type of density based outlier detection were employed in breast cancer based on point density [7]. The term point density denotes the remainder obtained from dividing the K-nearest neighbour with its k distance to detect the outlier.

The classical approaches is suitable for smaller datasets with limited number of dimensions. But for the larger datasets or high dimensional it faces a problem due to the distance and difference between the dataset is very small. This problem is overcome with the help of sub space learning, ensemble learning and clustering techniques [8]. It tested on the real time datasets like page blocks, arrhythmia, thyroid etc. the K-NN and Subspace based techniques is highly suitable for determine the outliers.

A two-step approach is proposed for identifying the outliers in high dimensional data in datasets like E.coli, spectrometer and Yeast [9]. The first step is to reduce the dimensions using Principal component analysis. The second step is to give the score and reduce the outlier using point density. The point density is estimated using kernel Density estimation process.

A combination of clustering and K-nearest neighbour is applied for identifying the outliers in the patient database [10]. Its main aim to improve the security and safety of the health care by removing the false records. By analysing various research works and papers, it is observed that the outlier detection technique is differ for each application and also it vary based on the size and attributes of dataset.

In this paper, an optimization based clustering algorithm is proposed for outlier detection for medical health care data. The K-Means clustering approach is highly suitable for medical dataset in determining the outlier. Its performance is further improved by using the dragon fly optimization. In medical field, the role of outlier detection is important in terms of predicting the disease at an early and storage space. Its performance is evaluated using detection rate.

The paper is organised by explaining the different contributions of outlier detection in medical and various fields in section2. Section 3 describes the existing method working and its drawbacks. The brief explanation of proposed method is given in section 4. The simulations and discussion of proposed method is explained in section 5. Section 6 conclude the work based on the results.

2. Literature Survey

An overview and comparison of three outlier detection techniques is discussed in [11]. The techniques used for the comparison are clustering based technique, density based and distance based technique. These techniques were analysed in terms of complexity, computational time, and accuracy of result and dimensionality of data. In all these four aspects, the clustering based technique is able to detect the outlier in all applications with minimal computational time at high efficiency irrespective of the size of data.

A combination clustering, density and K-Nearest neighbouring algorithm Pruning based K-Nearest Neighbour is used for the big medical health care data based outlier detection [12]. The larger dataset is divided into smaller datasets using attribute overlapping rate. Then, the smaller dataset is further processed based on classification Quality character. It helps to divide the larger datasets as smaller chunk and K-NN is applied to it to determine the outlier. The main drawback is combining the outcome to detect the perfect outlier.

A Grid based local density called GidlOF is used for detecting the outliers in the medical insurance dataset [13]. The accuracy of traditional LOF is improved by adding information entropy and the normal data is removed from the set which results in low computational time. It can also process in Hadoop and big data of medical insurance records. But, the method is highly designed only for medical insurance dataset.

A pattern based technique is proposed for outlier detection in medical field in [14]. First, the prefix tree is used to group the data. Then, the pattern is generated using FP-Growth technique. From the generated pattern, the rare patterns are used for the outlier detection. The outlier are detected based on the scores of three factors TOF, RPSDF and RPOF. This method is tested on the Wisconsin breast cancer dataset by increasing the records gradually. This process is suitable for lower records and dimensions.

An improved K-Means clustering is proposed for detecting the medical intruder in medical insurance dataset [15]. The K-means clustering is used to cluster the dataset and it uses the squared error and total variance to reduce the error rate. The outlier is detected when the point is having the multiple of distance between the neighbourhood points. It is tested on the medical datasets like heart disease, pneumonia and bronchitis. The main drawback is choosing the 'K' value for K-means clustering.

A brief explanation about the outlier techniques for the data streams is discussed in [16]. Data stream is differ from the static data outlier detection. Because, in static the whole dataset is modelled whereas in data stream it is not possible. The techniques for data streams are clustering, density based and sliding window. It also discussed about the challenges in outlier detection of data streams.

A different types of distance based outlier detection is proposed for the data streams in [17]. The DDOS technique is further classified into different types based on grouping and its size. The DDOS types are thresh _Leap, exact storm, approx.-storm, DUE, Abstract-C and MCOD. Apart from MCOD all other techniques were based on the sliding window process. The MCOD is based on the clustering process. Among these techniques,

the MCOD produced the best result. The drawback in this method it able to process the lesser number of attributes with higher records of dataset.

The major drawback in outlier detection is to determine the number of outlier points called top-n parameter. Most of the algorithms not able to solve this problem. This problem is solved by implementing the ROCF algorithm [18] using mutual neighbour graph method. In this, the number of outlier is the minimum number in the cluster point. The drawback in this approach it able to process only the lower attribute datasets like IRIS and not suitable for data streams and real time datasets.

A multi-objective based Genetic algorithm is used for detecting the outliers in Wisconsin Breast cancer dataset [19]. The outlier is determined based on the fitness function. The fitness function used in this work is the multi-objective. The first objective is based on the distance between the data points in the cluster. The second objective is the distance between the centroids of the neighbourhood clusters. The third objective is number of outliers in the solution. The outlier will be determined by minimizing all these objectives. This method has advantage by using multi-objective for predicting the accurate outlier as compared to the other methods. But, it is suitable for the lower records and attribute based dataset.

A new concept linguistic summary is used by the outlier detection in [20]. In this, the databases are sorted and the results are predicted in different categories. The predicted results are given as summary using linguistic manner. The predicted results are categorized in different groups with different limits. Based on the limits and group the linguistic output is displayed. It is tested on the diabetic dataset. It can be applied for smaller datasets but for the larger datasets, the query and computational time will increase.

3. Existing Method

The Existing 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 4.

3.1 Objectives of the proposed Method:

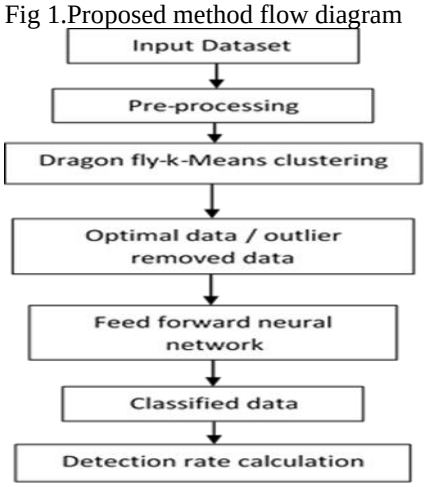
In this, the objectives of the Dragon fly K-means clustering is to overcome the drawbacks of the existing method. The other objectives of the proposed method are as follows:

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

4. Proposed Method

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 flow chart and explanation of the individual process is given below.

Fig 1.Proposed method flow diagram



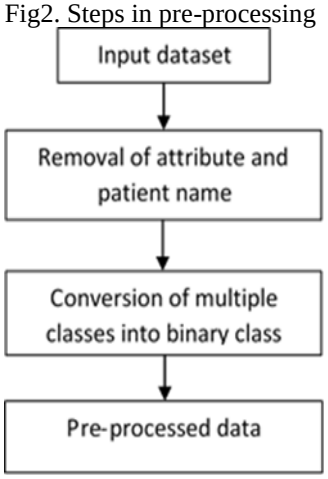
4.1 Input Dataset:

The proposed method is tested on the clinical datasets which requires high classification rate for detect the disease at an early stage. The clinical elliptical seizure dataset which is downloaded from the UCI machine learning repository database. The Iris flower dataset is also used for evaluating the proposed method performance

4.2 Pre-Processing:

The pre-processing step to convert the data into a suitable format to perform the proposed method process and to remove the irrelevant information from the dataset. The process in the pre-processing step is shown in the below figure.

Fig2. Steps in pre-processing

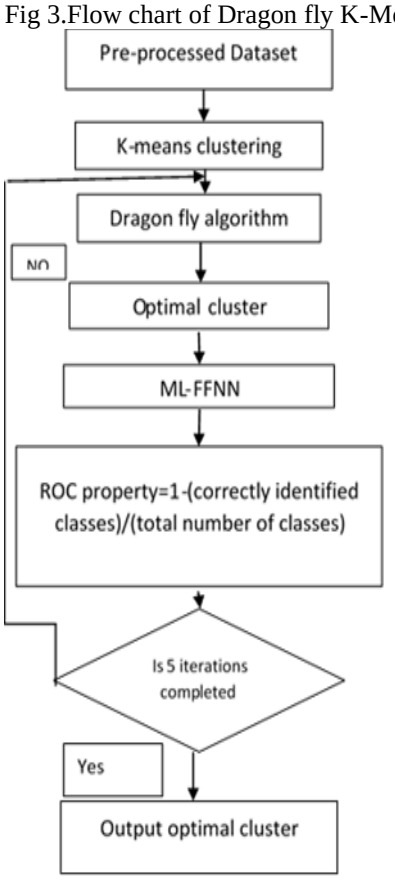


In this, the pre-processing steps is performed in two steps. The first step is to remove the attribute and patient name from the dataset. The second step is to convert the multi-classes into binary classes for easy outlier detection process. The multiple classes are converted into binary classes as normal and abnormal. These process are takes place on the dataset and the pre-processed dataset is used for the dragon fly K-means algorithm

4.3 Dragon fly K-Means

The pre-processed data is processed using dragon fly k-means clustering process to determine the optimal clustered data which removes the outlier data from the dataset. The working of dragon fly algorithm is inspired from [22]. But the objective function of the algorithm is the negative ratio property of the roc-curve of the classifier. The working of dragon fly algorithm is shown in below figure.

Fig 3.Flow chart of Dragon fly K-Means clustering



The steps in the dragon fly K-means algorithm is as follows:

Generally, all dragonflies are move towards the food source i.e., the objective or fitness function and distracted from the enemies. The main advantage of dragon fly algorithm is its ability to explore and exploit the search space. In this, the objective function is the reduction of negative rate of the classifier. The clustered data is given as input and the lower and upper boundary of the search space is 1 and 2. The five artificial dragon flies are used for searching the optimal clusters for five iterations till it minimize the ROC property. The phases for the dragonflies are explained below.

4.3.1 Separation phase:

In this phase, the collision between the current dragon fly (X) and neighbouring dragon fly (Xj) in the search space is avoided by the following equation 1.

$$Sj = -\sum_{j=1}^{5} X - Xj \qquad 1$$

1.1.1 Alignment phase:

In this, the velocity of the current dragon fly (X) and neighbouring dragon fly (Xj) in the search space is aligned together using the following equation 2.

$$Aj = \frac{\sum_{j=1}^{5} Vj}{5} \qquad 2$$

1.1.2 Cohesion Phase:

In this, the attraction of dragon fly toward the centre of the neighbourhood is calculated using the following equation 3.

$$Cj = \frac{\sum_{j=1}^{5} Xj}{5} - X \qquad 3$$

1.1.1 Attraction Phase:

In this, the attraction of dragon fly toward the food source that is the objective function of the algorithm is calculated using the following equation 4.

$$Ai = X^{+} - X \qquad 4$$

1.1.2 Distraction Phase:

In this, the dragon fly is distracted from the enemy is calculated using the following equation 5.

$$Di = X^{-} - X \qquad 5$$

1.1.3 Updating Position vector

After each iteration, the position of the dragon fly is updated using the following equation 6.

$$X_{t+1} = X_t + \Delta X_{t+1} \qquad 6$$

WhereΔX_{t+1} is the step vector and is calculated using following equation 7.

$$\Delta X_{t+1} = s(Sj) + a(Aj) + c(Cj) + a1(Ai) + d(Di) \qquad 7$$

Where s, a, c, a1 and d are the weights for the each phase to explore and exploit the search space.

To improve the position vector of the dragon flies, the levy flight calculation is added to it and the equation is given below

$$X_{t+1} = X_t + levy\ (X) \times X_t \qquad 8$$

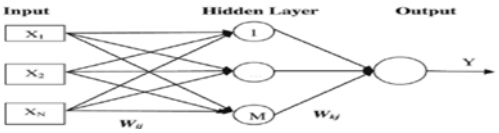
After each iteration, the local best solution, global best solution and position of the dragon fly is updated. Once, it reach the end of iteration the global best position is the optimal cluster and global best solution is the minimum ROC property. Then the optimum cluster will be used for training the network

1.2 Feed forward neural network (ML-FFNN)

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.

Fig 4. A schematic diagram of ML-FFNN



The input layer consists of the input weights based on the attributes which is denoted in terms of I neurons. The hidden layer neurons are denoted by j neurons. The connection between the input and hidden layer is made through the weight function $W_{ij\ which}$ is given by the following equation 9.

$$w_{ij}^{(K+1)} = W_{ij}^{K} - \lambda\left(\frac{\partial E}{\partial W_{ij}}\right)^{K} \quad (9)$$

Where $\lambda$the learning coefficient which is always greater than zero and K is indicates the output the layer.

The output Y is obtained from the hidden layer using the following equations

$$Y = f(\mathcal{E}_i) \quad (10)$$

Where$f(\mathcal{E}_i)$ the transfer function for propagating the signal from the j[th] neuron to i[th] neuron. The$(\mathcal{E}_i)$ is the potential function of the input neuron which is given by

$$\mathcal{E}_i = \vartheta_i \sum_{j\epsilon} w_{ij}.x_j \quad (11)$$

Where ϑ_i the threshold coefficient and W_{ij} is the weight coefficient which varies till it minimize the mean square error between the required output and predicted output through back propagation process.

The input of the feed forward neural network is the data from the optimum cluster and the hidden layer is 10 and the outputs are the targets of the data. The network is trained with the 70% of optimal data through cross-hold validation process and tested with the 30% of data. The classified output is evaluated using detection rate.

1.2 Outlier Detection:

In this, the outlier in the data is detected directly by applying the equation 12 to the column of the clustered data. The formula used to detect the outlier is as follows:

$$outlier = 1.4826 * median(abs(D) - median(D)) \quad (12)$$

In the above equation 12, the D represents clustered column data.

1.3 Evaluation metric

The outlier performance is evaluated using classification of datasets as normal and abnormal. The classification performance is evaluated using the detection or accuracy rate of the classifier as follows.

$$detection\ rate = \frac{correctly\ identified\ classes}{total\ number\ of\ classes} \quad (13)$$

Based on this, the proposed method performance is evaluated on three datasets and it is compared with the existing techniques. The results of the proposed method is explained elaborately in the following section.

2. Experimental Results and discussion:

The proposed method is simulated using Matrix Laboratory software R2018a version in windows 10 environment.

The proposed methodology is tested on the three types of openly available datasets namely Arrhythmia, diabetes and elliptical seizure datasets from the UCI machine learning repository site [24]. The description of the dataset is given below.

Tab1. Input dataset characteristics

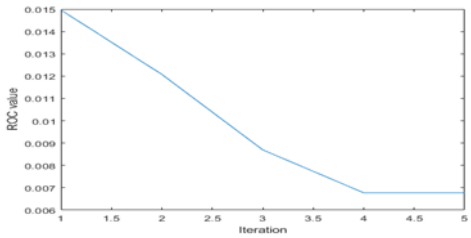
Parameter	**Elliptical seizure**	**IRIS flower**
No of records	11500	150

No of attributes	178	4
Actual classes	5	3
Converted classes	2	3
Existing techniques applied	No	No

Initially, the data is separated as inputs and targets for processing. Then the input data is clustered using K-means clustering process and from the cluster the optimal cluster is selected with the help of fitness function that is the Region of curve of the classifier.

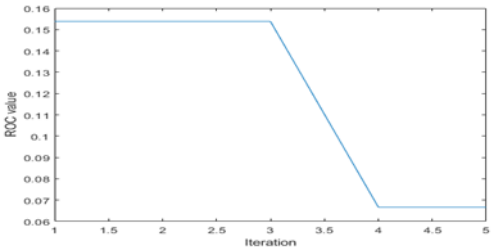
The convergence curve for selecting the optimal cluster for the classification for each disease is shown in the following figures

Fig 5. Elliptical seizure convergence curve



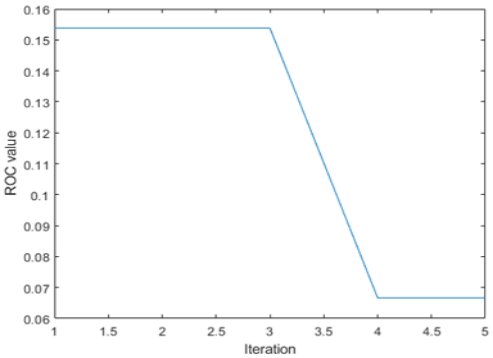
The above figure 5 shows that the proposed DF-K-Means for elliptical seizure datasets is converged after iteration 4 and the value is gradually decreasing from 0.015 to 0.0075 which is best for removing the anomaly data and produce the best classification result. The optimization time is high as compared to the other datasets due to the large number of records.

Fig 6. Sepal flower dataset convergence curve



From the above figure 6, the ROC value of the 0.155 is dropped to 0.065 after 5 iterations. From the figure, it is observed that the algorithm starts to converge after 3 and it maintain its stable position after 4th iteration.

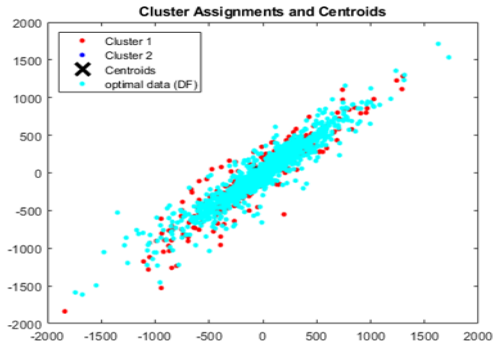
Fig 7. Petal dataset convergence curve



From the above figure 7, the ROC value of the 0.155 is dropped to 0.065 after 5 iterations. From the figure, it is observed that the algorithm starts to converge after 3 and it maintain its stable position after 4th iteration.

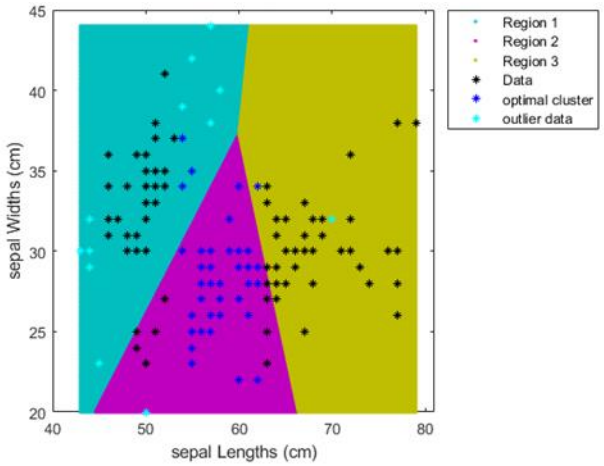
The dragon-fly based K-means clustering data is shown in the below figure.

Fig 8. Elliptical seizure DF_K-Means clustering



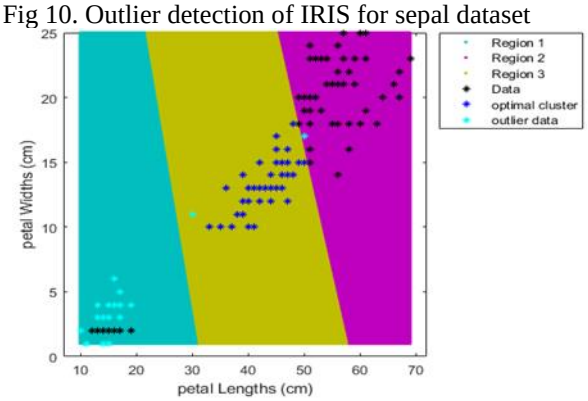
The above figure 8 is the overlaying of the optimal selected cluster points over the traditional k-means clustering of the data. The optimal attributes and cluster is defined by the cyan colour. The red colour shows that the cluster 1 data points.

Fig 9. Outlier detection of IRIS for sepal dataset



The figure 9 shows the three classes of the iris flower dataset in three colours and the data points are indicated in terms of the black colour. The optimal cluster points from dragon fly algorithm is mentioned in blue colour. The outlier points based on the proposed distance formula is mention in cyan colour. It also observed that the cluster points are more in the cluster region 1 as compared to the other clusters.

Fig 10. Outlier detection of IRIS for sepal dataset

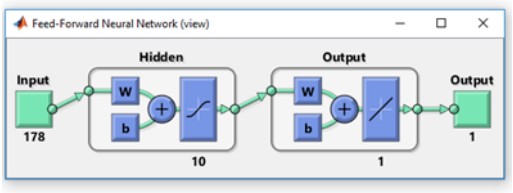


The figure 10 shows the petal data distribution of three classes of Iris flower dataset in three regions and black colour is used for indication of the data distribution. The data points are clustered using Dragon fly k-means to determine the optimal cluster and the optimal cluster points are shown in blue colour. The outlier data are indicated in cyan colour and it is calculated using the proposed distance formula.

From the Dragon fly based K-means clustering the optimal cluster for the classification is determined for individual datasets and it is trained using feed forward neural network. The optimal dataset is classified into

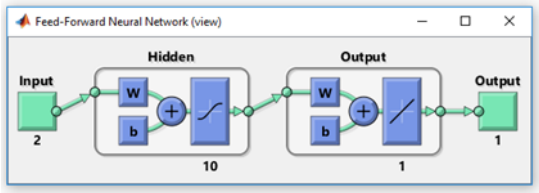
training and testing using hold out approach at 0.3%. The training data is used for training the network and testing data is used for testing and evaluation process.

Fig 11.Neural network diagram of Elliptical seizure



The above figure 11 is the training network for the elliptical seizure data set with 178 attributes as input and 10 hidden layers and an output layer.

Fig 12. Neural network for Iris data classification



The figure 12 shows the neural network diagram of the iris data classification. This diagram is common for both the petal and sepal classification. The input variable will be vary as petal or sepal based on the classification. Based on the input, the hidden and output neuron weight also varied.

The proposed DF-Means algorithm for outlier detection is tested on two datasets and the results of two datasets namely elliptical seizure and Iris dataset are calculated using the evaluation metric equation 13. The results are tabulated and it shown in the below table 2.

Tab 2. Comparison of outlier detection approaches

Dataset	**Proposed**
Elliptical seizure	0.9649
Iris dataset	0.9759

The above table 2 shows that the proposed method able to detect the outlier along with the classification of data with high detection rate of 0.9759 for the Iris data set and 0.9649 for the elliptical seizure dataset. It also indicates that the proposed DF-K-means able to process larger datasets like Elliptical seizure with 11500 records and smaller datasets like Iris flower with 150 records.

6. Conclusion:

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.

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.

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. Sheng, L. L. (2016). Fractal-based outlier detection algorithm over RFID data streams. International Journal of Online and Biomedical Engineering (iJOE), 12(01), 35-41.
7. Jha, G. K., Kumar, N., Ranjan, P., & Sharma, K. G. (2016). Density Based Outlier Detection (DBOD) in Data Mining: A Novel Approach. In Recent Advances in Mathematics, Statistics and Computer Science (pp. 403-412).
8. Xu, X., Liu, H., Li, L., & Yao, M. (2018). A comparison of outlier detection techniques for high-dimensional data. International Journal of Computational Intelligence Systems, 11(1), 652-662.
9. Kamalov, F., & Leung, H. H. (2020). Outlier detection in high dimensional data. Journal of Information & Knowledge Management, 19(01), 2040013.
10. Gebremeskel, G. B., Yi, C., He, Z., & Haile, D. (2016). Combined data mining techniques based patient data outlier detection for healthcare safety. International Journal of Intelligent Computing and Cybernetics.
11. Mandhare, H. C., & Idate, S. R. (2017, June). A comparative study of cluster based outlier detection, distance based outlier detection and density based outlier detection techniques. In 2017 International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 931-935). IEEE.
12. Yan, K., You, X., Ji, X., Yin, G., & Yang, F. (2016, October). A hybrid outlier detection method for health care big data. In 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)(BDCloud-SocialCom-SustainCom) (pp. 157-162). IEEE.
13. Xie, Z., Li, X., Wu, W., & Zhang, X. (2016, October). An improved outlier detection algorithm to medical insurance. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 436-445). Springer, Cham.
14. Borah, A., & Nath, B. (2019). Incremental rare pattern based approach for identifying outliers in medical data. Applied Soft Computing, 85, 105824.
15. WU, J., ZHANG, R., SHANG, X., & CHU, F. (2017). Medical insurance fraud recognition based on improved outlier detection algorithm. DEStech Transactions on Computer Science and Engineering, (aiea).
16. Thakkar, P., Vala, J., & Prajapati, V. (2016). Survey on outlier detection in data stream. Int. J. Comput. Appl, 136, 13-16.
17. Tran, L., Fan, L., & Shahabi, C. (2016). Distance-based outlier detection in data streams. Proceedings of the VLDB Endowment, 9(12), 1089-1100.
18. Huang, J., Zhu, Q., Yang, L., Cheng, D., & Wu, Q. (2017). A novel outlier cluster detection algorithm without top-n parameter. Knowledge-Based Systems, 121, 32-40.
19. Duraj, A., & Chomątek, Ł. (2017). Outlier detection using the multiobjective genetic algorithm. Journal of Applied Computer Science, 25(2), 29-42.
20. Duraj, A., Niewiadomski, A., & Szczepaniak, P. S. (2018). Outlier detection using linguistically quantified statements. International Journal of Intelligent Systems, 33(9), 1858-1868.
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.
22. Mirjalili, S. (2016). Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems. Neural Computing and Applications, 27(4), 1053-1073.
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.
24. https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition