Grenze International Journal of Engineering and Technology, Jan Issue

A Modified Fuzzy C-Means Algorithm using Fruit Fly Optimization for Data Clustering

S.Nagarjuna Reddy[1], S.Sai Satyanarayana Reddy[2] and M.Babu Reddy[3]
[1] Lakireddy Balireddy College of Engineering/Computer Science & Engineering, Mylavaram, India
Email: nagvik@gmail.com
[2]Sreays Institute of Engineering and Technology, Hyderabad, India
Email:saisn90@gmail.com
[3]Krishna University,MachiliPatnam, India
Email:m_babureddy@yahoo.com

***Abstract*— Clustering is the most promised technique used for data analysis. It is an unsupervised machine learning technique, means it does not require any target class. In the recent years so many clustering algorithms are invented with better accurate results. In this paper introduced a Modified Fuzzy C-Means Clustering algorithm based on Fruit-Fly optimization algorithm. Fruit-Fly optimization algorithm is used for extracting useful patterns from the data. Then our proposed approach will generate useful clusters. Our technique demonstrated to be an exceptionally effective method for the clustering of numerical data with high rate of accurateness. The function can be assessed in the name of sensitivity, specificity and accuracy. The accurateness can be associated with other prevailing methods.**

***Index Terms*— Unsupervised Learning, Clustering, Sensitivity, Specificity, Accuracy, Fuzzy C-Means Clustering.**

I. INTRODUCTION

Clustering is an unsupervised machine learning technique generates significant or valuable clusters that have related characteristics among all the data objects [1]. It is completely different from the supervised learning technique like classification, it contains target classes. Over the decade enormous clustering algorithms are developed to produce clusters with high accuracy [2].Clustering can be represented in two ways i.e. Hard Clustering and Fuzzy Clustering. If all the data objects are dedicated to one group then it will be treated as hard clustering. In case of Fuzzy clustering all data objects are not belonging to one group, instead they are going to share some fractional amount of membership in each cluster group [3]. In our proposed method we are using Fuzzy C-means (FCM) algorithm. It is one of the most useful fuzzy clustering methods. It generates accurate results when data contains less noise and highly sensitized to outliers. It also has few limitations while handling high dimensional data [4].

In cluster analysis, for many applications various graph-theoretic methods are used. To evaluate adjacent or neighborhood instances in the graph we require graph-based clustering techniques [5].

The main motivation behind our proposed work is to find the hidden structure among the data objects by grouping all the objects into clusters. Here we clustered the data objects using Modified Fuzzy C-Means Clustering method (MFCM). In our research we used Fruit fly optimization algorithm for data clustering [6]. It is used to extract useful patterns from the given data set.

Grenze ID: 01.GIJET.7.1.501_5

II. Literature Survey

Sai Satyanaryana et al. [7], have introduced the approach for organizing information objects into quality clusters. Our investigation mainly focused on Parkinson and Spam base data sets. Here, they used tree based strategy to found useful clusters with the help of binary cuckoo search. It produces better results compared to existing K-Means algorithm.

R. Senthil Kumar et al. [8] proposed an algorithm, based on the neural networks using back propagation techniques. They used various decision tree algorithms like M5 and REPTree for deriving optimal rules. They concluded fuzzy logic model gives better results compared to other computing techniques.

Musa Peker.[9] proposed an algorithm, to help medical professionals. He proposes an approach to solve the issues in the diagnosis phase itself. Mainly he focused on the three diseases. To analyze the performance they used various performance evolution measures like accuracy, specification and sensitivity, kappa statistics and ROC analysis were used.

Hazim iscan et al. [10] proposed an application of fruit fly optimization algorithm. They provided the solution for well known optimization problem i.e. travelling sales person problem using fruit fly optimization technique. In the modified approach the first phase follows the same approach and in the vision phase two different methods are developed. In the result the first phase will be updated based on the first 30 percent of the best solutions and the other part will be generated based on the random results.

Wen-TsaoPan[11] proposed an approach of finding the maximum and minimum value and it will be tested repeatedly. In this approach they the financial distress data of Taiwan's data collected. They adopted General Regression Neural Network, and multiple regressions to construct a model.

Ramdane Chikh et al. [12] proposed a new detection approach for e-mail based on the combined clustered NSA and fruit fly optimization algorithm (CNSA–FFO). In the proposed system they combine K-Means Clustering with FFO for increasing the efficiency of the traditional NSA. Results show CNSA-FFO produces better results than classic NSA.

Pan Wen-Tsao et al. [13] uses oil and gold data for his research. They are going to propose a cluster analysis algorithm to forecast the gold and oil prices.

Xiujuan Lei et al. [14] proposed a chaotic fruit fly optimization algorithm for finding the food behavior of fruit fly. To find the local optimum of Fruit fly optimization they proposed a new approach with Chaos operation. In the proposed approach to detect the movement of fruit fly they introduced logistic chaos mapping and going to generate local optimum for best fruit fly and best fly in Chaos.

III. Problem Identification

In the proposed approach we used an unsupervised machine learning method to extract information from the data sets i.e. clustering. To process the data using clustering we require unlabeled data. This is cheaper compared to labelled data. In data clustering, unsupervised machine learning method has some problems which are presented in existing papers that problems are discussed as follows.

- The unsupervised machine learning algorithm has large amount of data because of this reason speaking, designing of this algorithm is more difficult than supervised feature selection algorithms.
- In unsupervised learning, spectral classes don't correspond to information classes. The spectral properties of the classes can also change over time because of this reason user always use same class information when from one image to another image moving.
- In unsupervised learning, the goal is to discover similar group instants within the data. This method doesn't have static of data joint together with class labels of data.
- The unsupervised learning uses optimization criterion then partitioning of each node is done. This criterion is based on the object in the node only and which not to the other nodes splitting.
- In unsupervised learning the detecting features is hard task compared to selection of features in the supervised learning.
- The unsupervised learning method will fails when attributes in the feature are equal. It doesn't have scaling property invariance.
- In unsupervised learning method, the conditional distribution generates data sequences is strange. In that case it is hard to calculate required quantities precisely.

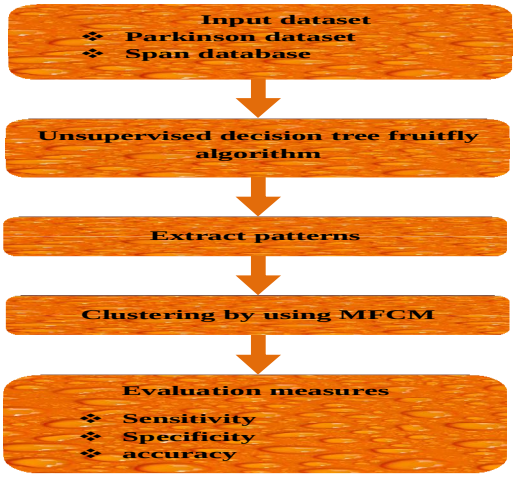


Fig 1: Proposed Architecture for data clustering

IV. PROPOSED METHODOLOGY

In our proposed research clustering is used to discover the hidden similar patterns from the given data set. Initially we propose a decision tree fuzzy min max neural network. In this investigation we examine with the help of a suited for prediction or forecasting needs. In this investigation we examine with the help of pattern based clustering method to clustering the small group of patterns. In the proposed research for extracting the patterns we introduced a Meta Heuristic algorithm named as fruit fly optimization algorithm. Fruit Fly algorithm is utilized to extract hidden patterns from in the hierarchical manner from the forest of unsupervised decision trees. In this phases can be performed with the aid of DT-FF technique, using this method to perform a patter extraction then the extracted patterns are feed in to the clustering phase. After pattern extraction using Improved Modified Fuzzy C-Means Clustering (MFCM) we can extract useful information in the form of clusters. We can adapt some pattern based clustering algorithms to extract patterns in our investigation.

Fruit Fly – Decision Tree Algorithm

In the proposed paper we used a Meta-Heuristic algorithm i.e., Decision tree – fruit fly algorithm for pattern extraction. The Fruit Fly Optimization Algorithm (FFOA) is another optimization strategy is used the behavior of the fruit fly to discover the food in optimized manner. In species category fruit fly is better in detection and recognition in terms of vision and olfactory. The fruit fly able to spot the food source from a distance of 40 Km and to notice all smells around in a particular range olfactory organ will be helpful. Based on the smell fruit fly flies in that direction by improvising the capability of its vision. The process made by fruit fly to finding the food can be summarized in various steps.

- Initially, Based on olfactory organ fruit fly smells the food source and flies in that direction.
- Based on the vision capability, fruit fly flies nearest location to the food.
- At last, other fruit flies also rush towards in that direction.

Steps of Proposed Approach:

Step 1: Initialization of required parameters. The proposed algorithm requires three main parameters to detect the location i.e. initial location of fruit fly, the size of population and total number of evolutions. Initialize the random location of a fruit fly by using the notations like (FU_{axis}, FV_{axis}).

Step 2: Traveling around using random path using olfactory organ and calculate the distance between initial location and detected path.

$$\mathrm{FU}_m = FU_{axis} + Rand \qquad (1)$$

$$\mathrm{FV}_m = FV_{axis} + Rand \qquad (2)$$

Step 3: Based on the obtained value compute the distance D, and the computed value of smell concentration ($\mathrm{S_C}$), known as the reverse of distance, is computed next.

$$D_m = \sqrt{(FU_m^2 + FV_m^2)} \quad (3)$$

$$S_{Cm} = 1/D_m \quad (4)$$

Step 4: By using fitness function, for every fruit fly position calculate the smell value.

$$Fitness = function(S_{Cm}) \quad (5)$$

Step 5: From all the locations find the best concentrated value i.e. nearest to the target location

$$[\text{Excellent Smell Index}] = Max(\text{function } (S_{Cm})) \quad (6)$$

Step 6: Retains the highest value and (FU, FV) coordinate, then the fruit fly flies in that direction. By using the vision organ the fruit fly will tremble in that direction.

$$\text{Smell Excellent} = \text{Excellent Smell} \quad (7)$$

$$FU\_axis = FU(Excellent\ index) \quad (8)$$

$$FV\_axis = FV(Excellent\ index) \quad (9)$$

Step 7: To optimize the process execute the steps 2 to 5 and then decide if the current smell value is larger than previous value. If true, process step 6 and calculate the weight using proposed approach.

Modified Fuzzy C-Means Clustering (Mfcm)

The proposed clustering algorithm follows same as FCM like repeated optimization because of the several iterations the existing approach frequently affected and not providing valuable conclusions for getting optimal solutions..

The Weighted Fuzzy Clustering algorithm can be represented in the following notation:

$$J(W,U,V) = \sum_{k=1}^{c}\sum_{i=1}^{n} u_{k,i}^2 (d_{k,i}^{(W)})^2 \quad (10)$$

Where $d_{k,i}^{(W)}$ is can found using the blow notation:

$$d_{k,i}^{(W)} = \sqrt{\sum_{j=1}^{d} W^{\beta} (x_{i,j} - v_{k,j})^2} \quad (11)$$

Minimizing the above equations as follows,

$$v_k = \sum_{i=1}^{n} (u_{k,i})^m x_i / \sum_{i=1}^{n} (u_{k,i})^m, \forall k = 1,.....c \quad (12)$$

$$u_{k,i} = 1/\sum_{k=1}^{c} (d_{k,i}^{(w)} / d_{k,\bar{k}}^{(w)})^{2/m-1} \quad (13)$$

The main steps of the proposed clustering algorithm as follows:

- Initialize the number of clusters C, and also initialize the matrix U for partition with random values.

$$\sum_{k=1}^{c} u_{k,j} = 1 \forall j \; and \; 0 < \sum_{i=1}^{n} u_{k,i} < n \forall k \quad (14)$$

- Based on equations (15) and (16), initialize the weighted Vector W with some random Values.
- By Using equation (12) finds the center value v_k.
- Then, Update the Matrix U with new values using (13)
- Finally update the weighted vector based on the obtained values.

$$W_j = \{0 \; if \; D_j = 0 \; and \; 1/\sum_{t=1}^{h} [D_j / D_t]^{1/B-1} \; if \; D_j \neq 0 \quad (15)$$

$$D_j = \sum_{k=1}^{k}\sum_{i=1}^{n} u_{k,i} (x_{i,j} - v_{k,j})^2 \quad (16)$$

- Repeat above steps until all requirements are satisfied for termination.

The acquired information can be clustered on the basis of Improved Modified Fuzzy C-Means Clustering (MFCM). In this investigation we examine with the help of a pattern based clustering method to clustering the small group of patterns. Our technique demonstrated to be an exceptionally effective method for the clustering of numerical data with high rate of accurateness. The function can be assessed in the name of sensitivity, specificity and accuracy.

V. Result and Discussion

To analyze the performance of the proposed approach we are going to construct the confusion matrix to find various performance measures like accuracy, sensitivity and sensitivity. To estimate the performance we are going to consider positive (TP, FP) and negative (TN, FN) from the confusion matrix.
The description of TP, TN, FP, and FN values for the clustering based assessment measures is quantified in the subsequent table I.

Table I: Parkinson Data Set Evolution Measures

	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity
Cluster 2	73	2	0	1	0.9868	0.9864	1
	120	9	1	1	0.9847	0.9917	0.9
Cluster 3	75	89	0	1	0.9939	0.9868	1
	91	85	1	1	0.9887	0.9891	0.9883
	95	99	1	2	0.9847	0.9793	0.99
Cluster4	58	63	1	1	0.9837	0.9830	0.9843
	38	48	2	1	0.9662	0.9743	0.96
	89	53	0	1	0.9930	0.9888	1
	75	63	1	0	0.9928	1	0.9843

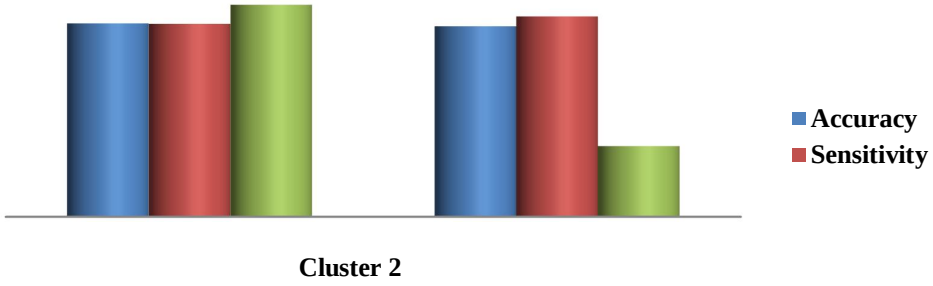


Fig 2: Parkinson Data set Evaluation Measures in Graphical Notation (Cluster-2)

Here we have discussed a evaluation measures for Parkinson dataset sensitivity measures. The above Table 1 and Fig 2 present all the evaluation measures. The specificity value is 0.958% for Cluster-2. The sensitivity value is 0.98% for Cluster-2. In our proposed research accuracy measures shows a very good performance of our research for Parkinson dataset for Cluster-2 attain accuracy measures such as 0.98%.

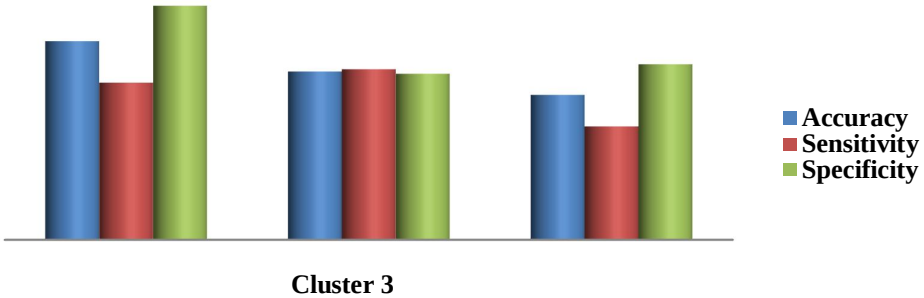


Fig 3: Parkinson Data set Evaluation Measures in Graphical Notation (Cluster-2)

Here we have discussed a evaluation measures for Parkinson dataset sensitivity measures. The above Table 1 and Fig 3 present all the evaluation measures. The specificity value is 0.99% for Cluster-3. The sensitivity value is 0.98% for Cluster-3. In our proposed research accuracy measures shows a very good performance of our research for Parkinson dataset for Cluster-3 attain accuracy measures such as 0.98%.

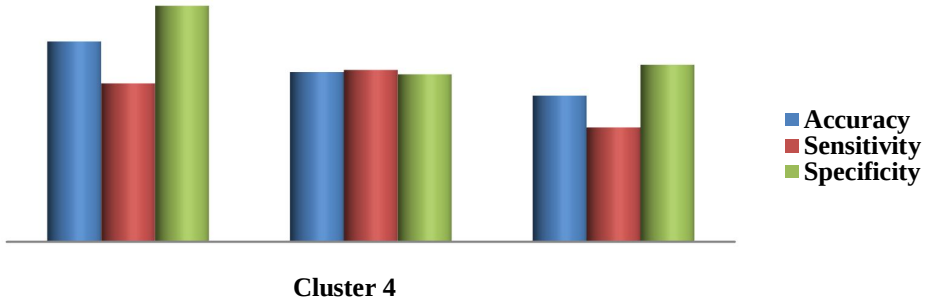


Fig 4: Parkinson Data set Evaluation Measures in Graphical Notation (Cluster-4)

TABLE II: SPAM BASE DATASET EVALUATION MEASURES

	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity
Cluster 2	341	2281	6	11	0.9935	0.9687	0.9973
	2507	481	8	16	0.9920	0.9936	0.9836
Cluster 3	708	10	2	11	0.9822	0.9847	0.8333
	953	21	10	8	0.9818	0.9916	0.6774
	953	255	10	8	0.9853	0.9916	0.9622
Cluster4	965	205	10	16	0.9782	0.9836	0.9534
	965	457	10	8	0.9875	0.9917	0.9785
	1108	508	7	8	0.9908	0.9928	0.9864
	857	508	18	8	0.9898	0.9907	0.9657

Here we have discussed a evaluation measures for Parkinson dataset sensitivity measures. The above Table 1 and Fig 4 present all the evaluation measures. The specificity value is 0.98% for Cluster-4. The sensitivity value is 0.98% for Cluster-4. In our proposed research accuracy measures shows a very good performance of our research for Parkinson dataset for Cluster-4 attain accuracy measures such as 0.98%.

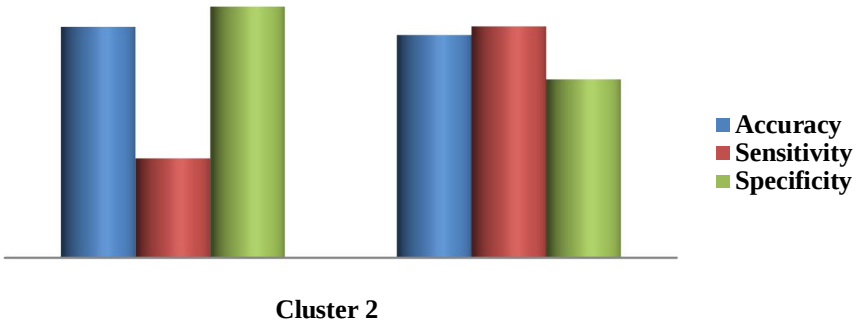
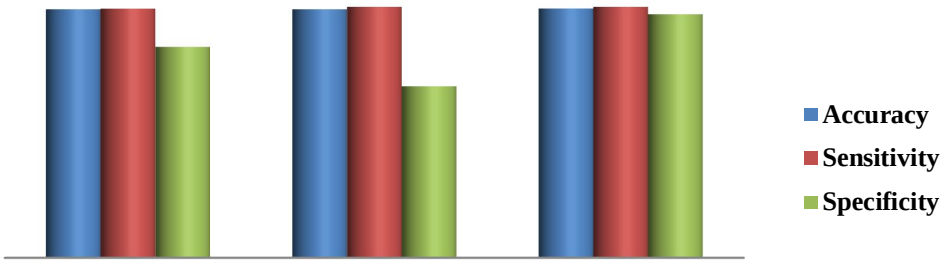


Fig 5: Spam Base Data set Evaluation Measures in Graphical Notation (Cluster-2)

Here we have discussed a evaluation measures for Spam Base dataset sensitivity measures. The above Table 2 and Fig 5 present all the evaluation measures. The specificity value is 0.99% for Cluster-2. The sensitivity value

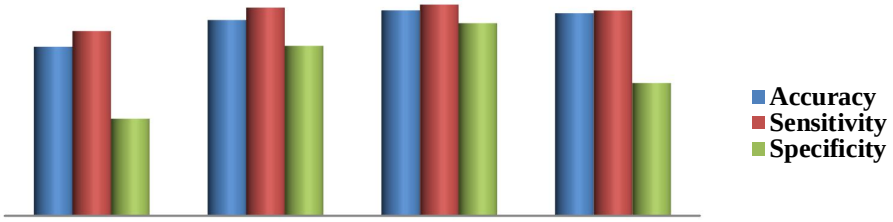
is 0.98% for Cluster-2. In our proposed research accuracy measures shows a very good performance of our research for Spam Base dataset for Cluster-2 attain accuracy measures such as 0.99%.



Cluster 3

Fig 6: Spam Base Data set Evaluation Measures in Graphical Notation (Cluster-3)

Here we have discussed a evaluation measures for Spam Base dataset sensitivity measures. The above Table 2 and Fig 6 present all the evaluation measures. The specificity value is 0.82% for Cluster-3. The sensitivity value is 0.98% for Cluster-3. In our proposed research accuracy measures shows a very good performance of our research for Spam Base dataset for Cluster-3 attain accuracy measures such as 0.98%.



Cluster 4

Fig 7: Spam Base Data set Evaluation Measures in Graphical Notation (Cluster-4)

Here we have discussed a evaluation measures for Spam Base dataset sensitivity measures. The above Table II and Fig 7 present all the evaluation measures. The specificity value is 0.97% for Cluster-3. The sensitivity value is 0.98% for Cluster-3. In our proposed research accuracy measures shows a very good performance of our research for Spam Base dataset for Cluster-3 attain accuracy measures such as 0.98%.

Analyze the Performance of Proposed Approach:

n this modified fuzzy c-means clustering based unsupervised decision tree-fruit fly based data clustering technique to be attained a very good accuracy for our proposed research. Also proposed approach gives better results while comparing with existing approach and results are presented in the below table. The assessment consequences are accessible in the subsequent Table III.

TABLE III: ASSESSMENT FOR PROJECTED AND PREVAILING ACCURACY RESULTS

Clusters	Proposed MFCM		Existing K-Means Clustering	
	Accuracy for Parkinson dataset	Accuracy for Spam Dataset	Accuracy for Parkinson dataset	Accuracy for Spam Dataset
2	98%	99%	98%	97%
3	98%	98%	97%	97%
4	98%	98%	98%	98%

The modified FCM results show the efficiency of our proposed research method would produce or correlate better accuracy values compared to the existing algorithms. From the produced consequences, our suggested modified FCM clustering gives better results in all performance measures.

VI. Conclusion

In the proposed paper we discussed a novel technique for clustering the data using a Meta heuristic algorithm like fruit fly optimization. The accuracy for the existed K-Means clustering using Binary Cuckoo Search is about 98% for the selected data set like Parkinson. Whereas, in the proposed approach also gives the same accuracy with slight improvements. But in case of spam mail data set proposed approach gives more accuracy compared to the existing approach. In Future we will try to use this approach for high dimensional data with various variant algorithms in the deep learning.

References

[1] K.Murugan, P.Varalakshmi, R.Nandha Kumar and S.Boobalan, "Data Mining Using Integration Of Clustering And Decision Tree," Volume-1, Issue -2, pp. 2347 - 2812, 2013.

[2] Vanisri, D, "A Novel Kernel Based Fuzzy C Means Clustering With Cluster Validity Measures," International journal of computer science and network solutions, vol. 2, no. 12, pp. 254-260, 2014.

[3] Gosain, Anjana, and Sonika Dahiya, "Performance Analysis of Various Fuzzy Clustering Algorithms." Procedia Computer Science, vol. 79, pp. 100-111, 2016.

[4] Pradeep Kumar, "Fuzzy based clustering algorithm for privacy preserving data mining" ,International Journal of Business Information Systems, Vol. 7, No. 1, pp.27-40, 2011.

[5] D. Karthika and S. Divakar, "Improving the Efficiency he Efficiency of Fast Using Semantic", International Journal of Scientific and Research Publication, Vol.4, Issue.1, pp. 1-5, Jan 2014.

[6] Chenping Hou, Feiping Nie, Xuelong Li, Dongyun Yi, and Yi Wu, "Joint Embedding Learning and Sparse Regression: A Framework for Unsupervised Feature Selection," IEEE Transactions on Cybernetics, Vol. 44, NO. 6, pp. 793-804, 2014.

[7] Sai Satyanaryana Reddy Seelam, Babu Reddy Mukkala and Nagarjuna Reddy Seelam, " Optimal Decision Tree Based Unsupervised Learning Method for Data Clustering", International Journal of intelligent Engineering and systems, no.21, pp. 126-135, 2016.

[8] R. Senthil kumar and Manish Kumar Goyal, "Application of ANN, Fuzzy Logic and Decision Tree Algorithms for the Development of Reservoir Operating Rules", Water Resource Manage, vol. 27, pp. 911–925, 2013.

[9] Musa Peker, "A decision support system to improve medical diagnosis using a combination of k-medoids clustering based attribute weighting and SVM", system level quality improvement, vol.40, pp.1-12, 2016.

[10] Hazim iscan, Mesut Gunduz," An Application of fruit fly optimization algorithm for travelling sales person problem",Advances in Information Technology, Volume 111, Pages 58-63,2017.

[11] Wen-TsaoPan," A new Fruit Fly Optimization Algorithm: Taking the financial distress model as an example", Knowledge-Based Systems, Volume 26, Pages 69-74, 2012.

[12] Ramdane Chikh,Salim Chikhi," Clustered negative selection algorithm and fruit fly optimization for email spam detection", Journal of Ambient Intelligence and Humanized Computing volume 10, pages143–152(2019).

[13] Pan, W.-T. (2014), "Mixed modified fruit fly optimization algorithm with general regression neural network to build oil and gold prices forecasting model", Kybernetes, Vol. 43 No. 7, pp. 1053-1063. https://doi.org/10.1108/K-02-2014-0024.

[14] Qinbao Song, Jingjie Ni, and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 25, no 1, pp. 1-13, 2013.