IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-8, Aug 2018 E-ISSN: 2347-2693

Min-Max based K-means Clustering Algorithm using Artificial Neural Network Approach

Gurpreet Virdi[1*], Neena Madan[2]

[1]Department of Computer Science and Engineering, Guru Nanak Dev University Regional Campus, Jalandhar, India
[2]Department of Computer Science and Engineering, Guru Nanak Dev University Regional Campus, Jalandhar, India

[*]*Corresponding Author: gvirdi47@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 18/Aug/2018, Published: 31/Aug/2018

Abstract— K-means clustering approach is the most commonly used approach to reduce the sum of intra-cluster differences. But there is problem regarding the selection of centroid in k means clustering algorithm. Centroid can be poor or best depending upon the data. Therefore, there is a probablility of selecting good or bad centroid. So, in case of poor centroid selection, data does not get clustered in proper manners. To overcome this problem, we have used Min-max based K-means clustering algorithm along with ANN (Minimum- maximum based artificial neural network). The ANN algorithm overcomes the pitfalls of Min-max based K-means algorithm (Poor selection of centroid). In our research we have used Min-max K-means algorithm along with ANN to find out the exact category according to the labeled input data. Here, ANN is firstly trained with labeled input data. On the basis of training, testing phase is done to determine the accurate output for labeled input data. The enhancement in the accuracy of the proposed work from the existing work is approximately 16.47%.

Keywords— Clustering, K-mean, Min-Max, ANN

I. INTRODUCTION

Clustering is the most essential unsupervised learning problem which is found in pattern identification, image processing and machine learning techniques [1]. The goal of clustering is to integrate the similar data into one group. An example of clustering is defined below.

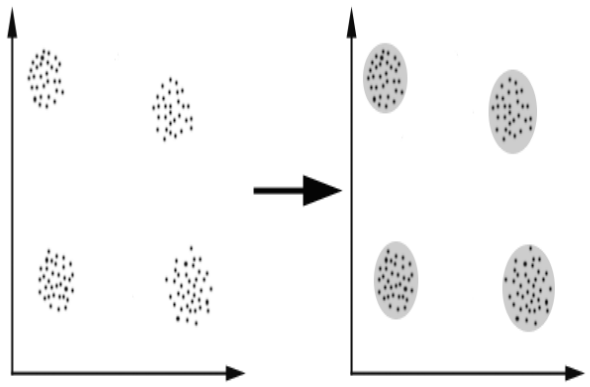

Figure1: Example of Clustering

Source:https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/

In the above figure, we are taking four groups of data that has been converted into clusters after applying clustering algorithm [2]. The clustering is performed on the basis of distance. According to the Euclidean distance formula, the objects that are close to each other come in a similar group. The formula is written below.

$$d(x,y) = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+(x_3-y_3)^2+(x_4-y_4)^2} \quad (1)$$

Here x_1, y_1 is the position of the 1st centroid, x_2, y_2 and x_3, y_3 are the position of the 2nd and 3rd centroid respectively; x_4 and y_4 are the position of the 4th centroid[3].

The main aim of clustering is to position the unbalanced data into balanced data. But how one can select an appropriate algorithm that creates a good clustering? In previous research, there are number of methods that have been performed to achieve better clustering like K-means clustering, Fuzzy c-means, hierarchical and mixture of Gaussians clustering [4]. Among all of the clustering techniques, K-means is the best clustering approach, in which clusters are formed by reducing the clustering error [5]. In K-means clustering algorithm, the data is classified through a certain number of K-cluster. In this clustering approach, the centroid is created using mean value of dissimilar group. On the basis of similarity index and K-

means, the data is divided into dissimilar clusters [6, 7]. In this research work, we have adopted the K-means approach and has integrated Min-max algorithm along with ANN approach to enhance the efficiency of the system.

The paper consists of various sections. Section I contains the introduction about the research topic. Section II contains materials and methods that are used in research. In this we discussed Min-max k means and ANN. Section III contains experiment and analysis that is in this we had discussed about the experiment we did in matlab and showed the improved results through bar graphs. Section IV contains conclusion and future scope. In this section we discussed the conclusion that is what we had concluded and what we can do in future for further improvements.

II. MATERIAL AND METHODS

In this paper, we have used Min-max algorithm based on K-means algorithm using ANN as a classification technique. The brief description of Min-max along with K-means and Min-max with ANN is defined in section *A* and *B* respectively.

A. Min-Max K-means

As we know that, K means algorithm is used to decrease the sum of intra cluster difference[8]. The problem with K-means clustering is that, it can cause a bad formation of centroid. To overcome this problem of K-means algorithm, Min-Max K means approach has been used by author (Tzortzis et al., 2014)[9]. In this approach, the weights are assigned to the cluster with respect to their difference and the weighted value of the K-means to optimize.

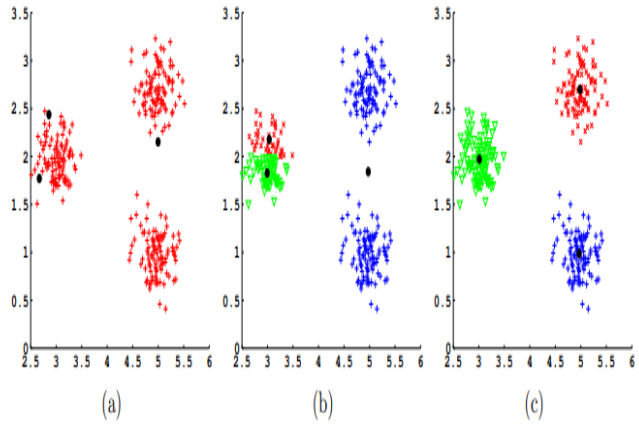

Figure 2: data clustering (a) Centroid initialized (b) K-means clustering (c) Min-Max K-means [9]

The above figure represents the unstructured data with centroid. The centroid assign to the cluster may be good or bad. But in the above case the centroid is bad as shown in figure 2 (a). For clustering the unstructured data, K-means algorithm is applied as shown in figure 2(b). As the centroid is bad; therefore; the clustering formed using K-means algorithm is not good. Thus, to enhance the clustering of K-means algorithm, Min-max with K-mean is applied and the structure formed is shown in figure 2 (c). In this figure, the data is clustered with best centroid. The algorithm for Min-max K-means is defined below.

Algorithm1: Min-Max K-mean

Input: Unstructured Data
Output: Clustered data

Initialize Min-Max with parameters –
Minimum Value (-Inf)
Maximum Value (+Inf)
For each Unstructured Data
Repeat until all data not simulated
Apply Min function = max (-Inf, Data)
Apply Max function = min (+Inf, Data)
Best cluster data = Comparison (Min function, Max function)
End
Return; clustered data
End

The drawback of Min-max K means algorithm is that the execution time is more as it compares the individual data within the group or wrong clustering is formed. To overcome this problem, ANN algorithm is used.

B. ANN (Artificial Neural Network)

ANN is a machine learning algorithm used to classify the accurate data within the appropriate cluster according to labeled input data [10]. In traditional approach, the complexity to find the appropriate cluster is more as the approaches that require more time [11]. To overcome this problem, ANN is trained with input data according to labeled data mainly categorized into three categories and on the basis of training of ANN, we can find the best cluster for the data [12]. The algorithm of ANN used in the proposed work is written below:

Algorithm 2: ANN

Input: Labelled data as a Training Data (T),
Target as a label (Gp) and Neurons (Ns)
Output: cluster data

Initialize ANN with parameters
– Iteration (I)
– Neurons (Ns)
– Performance metrics: MSE, Gradient, Mutation (Mu) and Validation check
– Training algorithm: Levenberg Marquardt (Trainlm)

– Data distribution: Random
For each set of T
Target = Object types of Training data
End
Initialize ANN using Training data and Group
Net = Newff (*T*, G_p, N_s)
Set the training parameters as per the necessities and train the system
Net = Train (*Net, Training data, Group*)
Return; cluster data
End

The training algorithm of proposed research work is explained below.

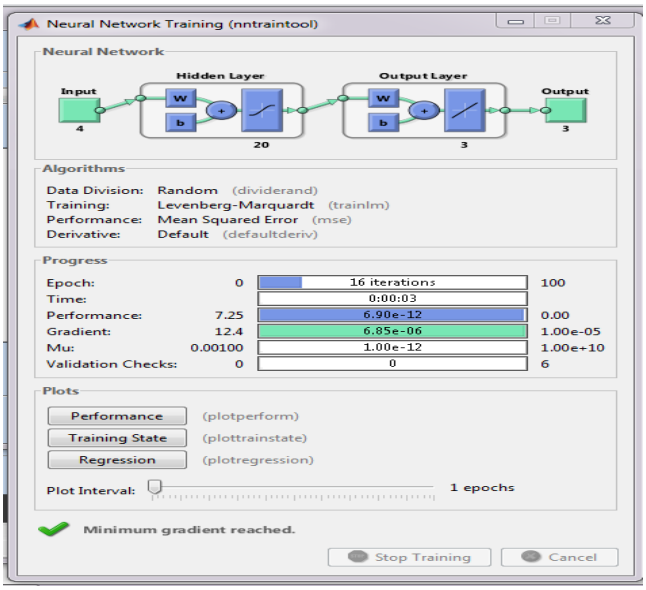

Figure 3: Architecture of trained ANN

The above figure describes the training process of ANN architecture. The ANN consists of three layers named as input layer, hidden layer and output layer. There are four neurons that are provided at the input layer of the ANN. The input neurons are passed to the hidden layer in which weights and bias is added to obtain the desired output. In this research work, the number of neurons in the hidden are 20 and in the output we have obtained 3 neurons. The maximum value of epoch used is 100 whereas the ANN is trained at 16 iterations [13].

III. EXPERIMENT AND ANALYSIS

The experiment has been performed on "Fisher Iris" data set that comprises of 150 instances , four attributes and 3 number of classes named as Iris setosa, iris versicolour and iris virginica. The experiment has been simulated in MATLAB simulator tool using Min-max based k-means algorithm using ANN algorithm. The performance parameters observed for both the algorithms are discussed below:

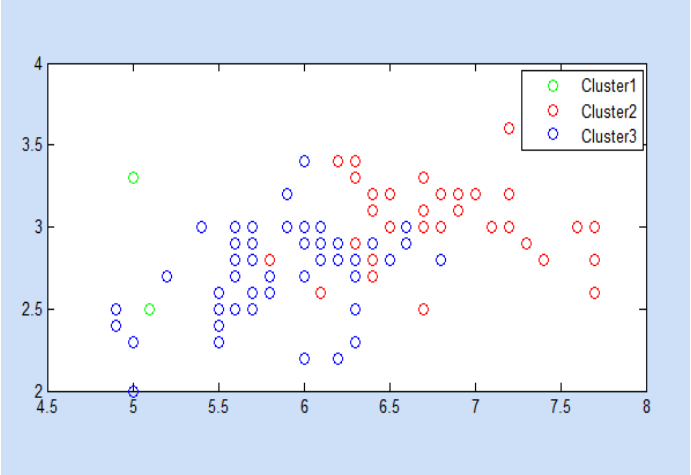

Figure 4: Cluster formation using Min-Max k means

The above figure defines the formation of cluster using Min-Max K-means algorithm. The clusters are formed in three groups represented by three different colours red, green and blue. Due to bad centroid selection the clustering of data is not proper which is given in above figure with three types of cluster. After applying Min-Max based K-means algorithm, the clustering is formed using Euclidean distance formula with an error of 2.2361. From the experiment, it has been observed that using Min-Max k-means algorithm, there are 11 numbers of objects from different clusters exist in cluster 1, in cluster 2 there is no incorrect data element and in cluster 3, there is 1 element of another group.

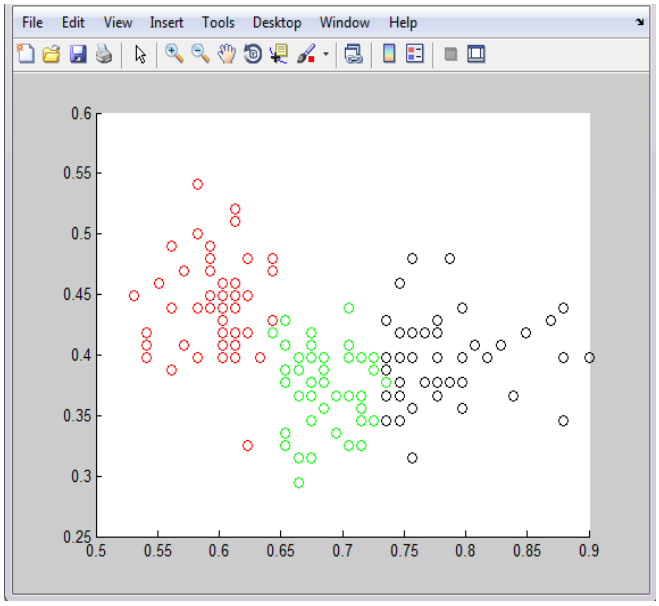

Figure 5: Cluster formation by Min-Max-ANN

The above figure defines the formation of cluster using Min Max based ANN algorithm. When ANN machine learning approach is used with Min-Max algorithm, the problem of complexity along with mixing of data within cluster formation is overcome. The result with best centroid is shown in figure 5. The cluster error is reduced from 2.2361 to 1.4142 when ANN is applied.

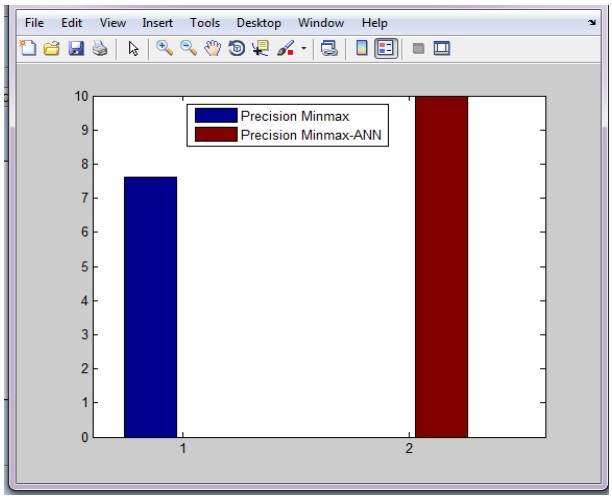

Figure 6: Precision comparison between existed and proposed work

The above figure defines the precision of proposed work with existing work. X-axis defines the type of algorithm and Y-axis defines the precision values. Blue bar line and brown bar line represents the value of precision for Min-max k means and Min-max based ANN algorithm. From the above graph, it is clear that the precision of Min-max based ANN algorithm increased from 7.8 to 10. Thus, there is an increment of 28.20 %.

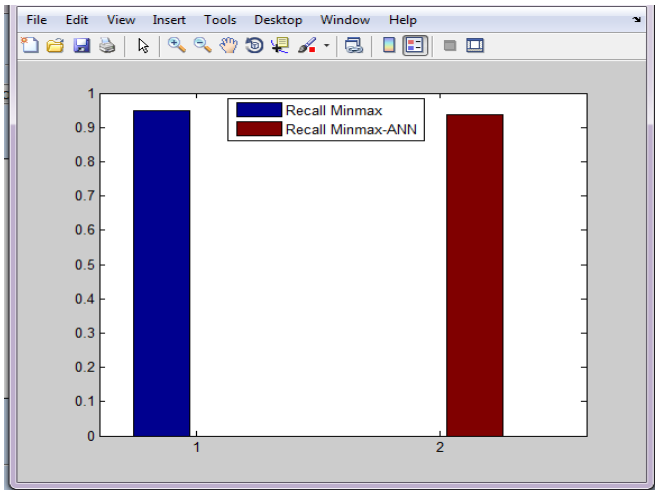

Figure 7: Recall comparison between existed and proposed work

Above figure displays the comparison of Recall values observed for Min-max k means and Min-max based ANN algorithm. The values of Recall observed for Min-max k means and Min-max based ANN algorithm are 0.95 and 0.9 respectively. From the above graph, it is clear that the recall value of proposed work is less as compared to the existing work. The percentage decrease in the Recall value from existing work to proposed work is 5.26%.

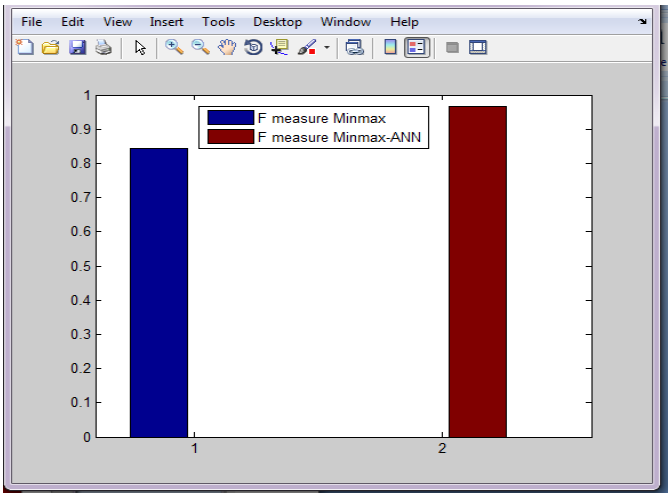

Figure 8: F measure comparison between existed and proposed work

The above figure displays the graph comparison between F-measure values for Min-max k means and Min-max ANN. Blue bar and brown bar represents the F- measure values for Min-max k means and Min max with ANN algorithm. From the above graph, it is observed that the F-measure value for the proposed algorithm is increased by 19.51%.

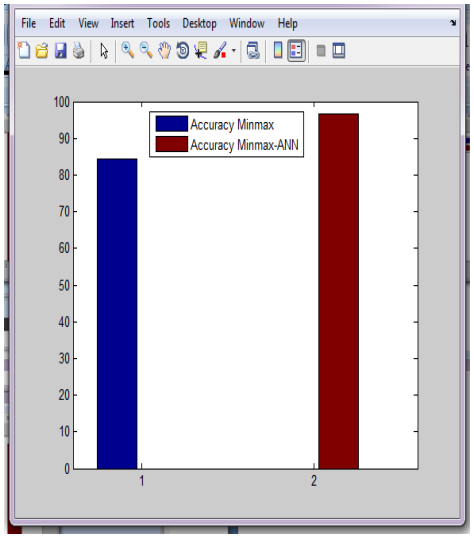

Figure 9: Accuracy comparison between existed and proposed work

The figure 9 represents the value of accuracy observed for existing work (Min-max k means) and for proposed work (Min-max based ANN). The increased in the accuracy of the proposed work from the existing work is approximately 16.47%.

IV. CONCLUSION AND FUTURE SCOPE

The accuracy of existing Min-max based K-means algorithm is enhanced in this proposed work using Min-max K-means based on ANN approach. The experiment has been performed using Min-max based K-means algorithm and Min-max based ANN approach to evaluate the performance on the basis of clustering error rate, number of objects that are incorrectly clusters and the performance parameters such as precision, recall, F-measure and accuracy. From the simulation results, we have observed that the proposed approach has performed well with high accuracy up to 99 %.

In future work, we can use any optimization algorithm such as swarm optimization to obtain the optimized cluster. And the training of ANN is performed on the optimized features. This helps to increase the efficiency of the proposed work.

REFERENCES

[1] A. Chadha, "*Efficient Clustering Algorithms in Educational Data Mining"*, Handbook of Research on Knowledge Management for Contemporary Business Environments (pp. 279-312). IGI Global, 2018.

[2] M. Kalra, N. Lal, & S. Qamar, "*K-Mean Clustering Algorithm Approach for Data Mining of Heterogeneous Data*", Information and Communication Technology for Sustainable Development (pp. 61-70). Springer, Singapore, 2018.

[3] Juntao Wang, Xiaolong, "*An improved k means clustering algorithm",* IEEE 3rd International Conference on Communication Software and Networks, 2018.

[4] A. Bansal, M. Sharma & S. Goel, "*Improved k-mean clustering algorithm for prediction analysis using classification technique in data mining*", International Journal of Computer Applications (0975–8887), Volume 157, 33-40, 2017.

[5] K. Vaswani, & A. M. Karandikar, , "*An Algorithm for Spatial Data Mining using Clustering*", Journal of Engineering and Applied Sciences,2017

[6] K.Teknomo, "*K-means clustering tutorial*", Medicine, 100(4), 3, 2006.

[7] N. K.Visalakshi, & J. Suguna, "*K-means clustering using Max-min distance measure*", Fuzzy Information Processing Society, 2009. NAFIPS 2009, Annual Meeting of the North American (pp. 1-6). IEEE, 2009.

[8] M. K. Yadav, & S. Sharma, *"A SURVEY OF FAST AND EFFICIENT K MEANS CLUSTERING ALGORITHM*", International Journal of Engineering, Management & Medical Research (IJEMMR), Vol 1, no. 9, 2015.

[9] G. Tzortzis, & A. Likas, "*The MinMax k-Means clustering algorithm*", Pattern Recognition, 47(7), 2505-2516, 2014.

[10] D. K. Ghosh & S. Ari, "*A static hand gesture recognition algorithm using k-mean based radial basis function neural network*", Information, Communications and Signal Processing (ICICS) 2011 8th International Conference on (pp. 1-5). IEEE, 2011.

[11] R. J. Schalkoff, "*Artificial neural networks*", New York: McGraw-Hill, 2011.

[12] B. Yegnanarayana, "*Artificial neural networks*", PHI Learning Pvt. Ltd, 2011.

[13] Z. Zhang, "*Artificial neural network",* Multivariate Time Series Analysis in Climate and Environmental Research (pp. 1-35). Springer, Cham, 2018.

