Comparison of Clustering with Self Organizing Map and Fuzzy C-Means Algorithm

Hsu Mon Maung, Tha Pyay Win
Computer University (Taunggyi)
hsumon77@gmail.com, thapyay14@gmail.com

Abstract

Clustering partitions a set of objects into non-overlapping subsets called clusters such that objects inside each cluster are similar to each other and objects from different clusters are not similar. The set of non-overlapping clusters is called a partition. Neural networks are believed to possess some particularly valuable properties, since they are patterned after associative neural properties of the brain. Neural networks proceed by a process called learning. The Self-Organizing Map (SOM) is a stable neural network model for high-dimensional data analysis. Most classical clustering algorithms assign each data to exactly one cluster, thus forming a crisp partition of the given data, but fuzzy clustering allows for degrees of membership, to which data belongs to different clusters. The best known fuzzy clustering algorithm is fuzzy c-means (FCM) clustering algorithm which is straightforward generalization of classical crisp c-means algorithm. This system is implemented clustering multidimensional data by using SOM and FCM algorithms.

1. Introduction

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding *a* structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. Data clustering is under vigorous development contribution areas of research include data mining, statistics, machine learning, spatial database technology, biology, and marketing. Clustering is an example of unsupervised learning. Unlike classification, clustering and unsupervised learning do not rely on predefined classes and class-labeled training examples. For this reason, clustering is a form of learning by observation, rather than learning by examples [7].

So it is implemented in clustering by using FCM and SOM algorithm to analyze the output results with accuracy measure. The rest of the paper is described as follows: section 2 presents the system overview. Section 3 is explanation of SOM algorithm and section 4 is fulfilled with FCM algorithm. Section 5 is included system design and implementation. Section 6 shows the experimental results of the system and the compared accuracy of the two algorithms. The conclusion of this system is combined at the last section.

2. System Overview

First, a set of input item with their attributes (**X**) are selected from database. Each input data (x_i) is an m-dimensional vector of m elements or m features. Since the m features in each input can have different units. In a geometric sense, each x_i is a point in m-dimensional feature space, and the set of data (**X**) is a point set with n elements.

A neural network's ability to perform computations is based on the hope that human can reproduce some of the flexibility and power of the human brain by artificial means. Neural networks also differ from each other in their learning rules that establish when and how the connecting weights change. Finally, networks exhibit different speeds and efficiency of learning. As a result, they also differ in their ability to accurately respond to the presented at the input. In contrast to conventional computers, which are programmed to perform specific tasks, most neural networks must be taught, or trained. They learn new functional dependencies. Learning corresponds to parameter changes. Learning rules and algorithms used for experiential training of networks replace the programming required for conventional computation. The Self-Organizing Map (SOM) is a stable neural network model for high- dimensional data analysis. However, its applicability is limited by the fact that some knowledge about the data is required to define the size of the network[10].

In fuzzy cluster analysis the Fuzzy c-Means (FCM) algorithm computes clusters centers'

coordinates and the partition matrix from the specification of the number of clusters that must be given in advance. In this system, the FCM algorithm is executed to various values of clusters and results are evaluated. This system demonstrates the benefits of this neural network model and fuzzy clustering algorithm using example from the zoo dataset clustering domain and other data set can also be used and analyzed.

3. SOM Algorithm

1. Randomly initialise all weights
2. Select input vector $x = [x_1, x_2, x_3, \dots, x_n]$
3. Compare x with weights w_j for each neuron j to determine winner
4. Update winner so that it becomes more like x, together with the winner's neighbours
5. Repeat from (2) until the map has converged (i.e. no noticeable changes in the weights).

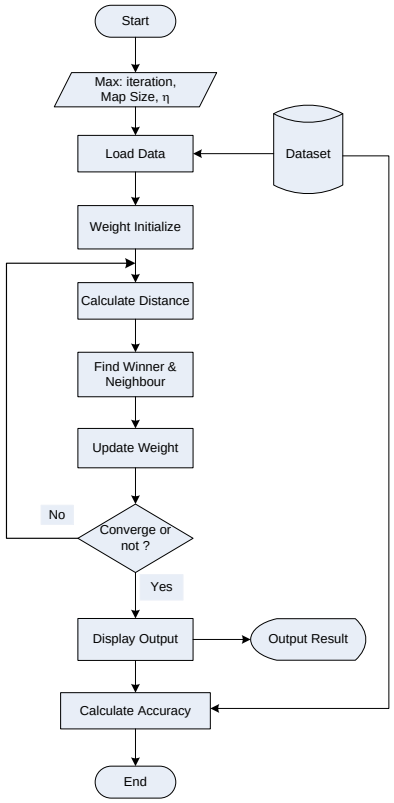

Figure 1. Process with SOM algorithm

4. FCM Algorithm

1. The algorithm is implemented by initializing the value of how many input data and clusters → partition matrix U(0) ,

m' → Fuzziness (1.2 - 2), and

ε_L → Accuracy (0.01 - 0.3) coefficients.

2. C centers are calculated for **r**th step.

$$v_{(L)}^{(r)} = \frac{\sum_{u} \mu_{u,}^{m'} \cdot x^{(r)}}{\sum_{u} \mu_{u,}^{m'} \cdot x^{(r)}}$$

3. Partition matrix for the **r**th step is updated:

$$\mu_{ik}^{(r+1)} = \left[\sum_{j=1}^{c} \left(\frac{d_{ik}^{(r)}}{d_{jk}^{(r)}} \right)^{2/(m'-1)} \right]^{-1}$$

4. The error is compared against the set value and the procedure is repeated if the error is not lesser than the set value

If $\left\| U^{(r+1)} - U^{(r)} \right\| \leq \in_L$ stop; else increment r by 1 and repeat the procedure from step 2.

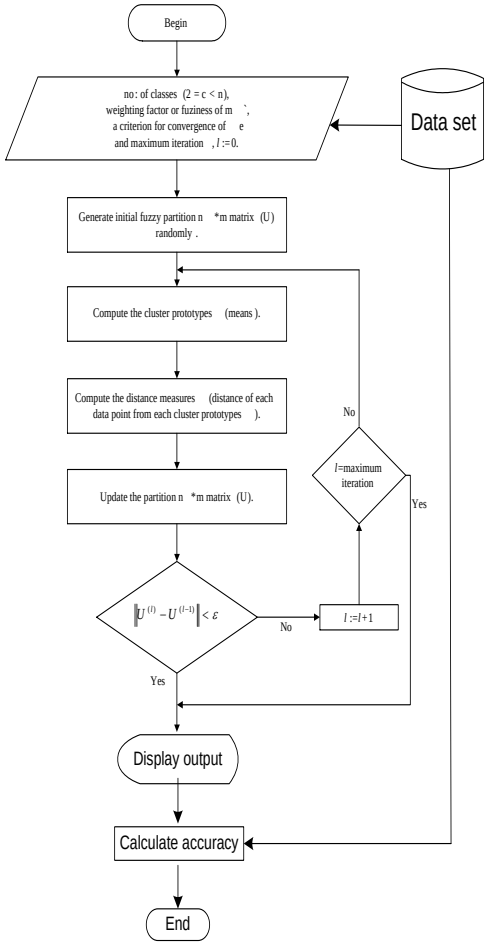

Figure 2. Flow of FCM algorithm

Fuzzy clustering method allows the objects to belong to several clusters simultaneously, with different degrees of membership. In many situations, fuzzy clustering is more natural than hard clustering. Objects on the boundaries between several classes are not forced to fully belong to one

of the classes, but rather are assigned membership degrees between 0 and 1 indicating their partial membership[4].

5. System Design

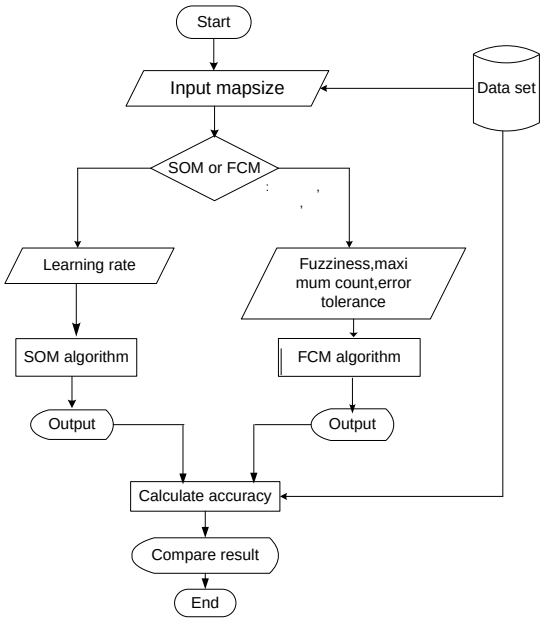

Figure 3. System flow diagram

This system also resolved implementation issues related to automatic clustering technique of Kohonen's Self-Organizing Map and Fuzzy c-mean algorithm. Firstly user can set the clustered matrix size and then user can choose the clustering method of either SOM clustering and FCM clustering, if SOM clustering is chosen, user must input learning rate. After then the system will process SOM clustering and give cluster result with given matrix size. If FCM clustering is chosen, user must input fuzziness or weighting factor (m'), maximum loop count, and error tolerance. After then the system will process according to FCM clustering and finally show cluster result. The system also calculate the accuracy of the two algorithms. Accuracy is calculated by comparing with given dataset. After calculating accuracies of two algorithms, the system show the compare result. The user can view and study the clustering result of partitioning based clustering algorithms and compare the result with manual and machine learning based results. In this system, the clustering results can be demonstrated with map sizes of up to 5x5 matrices.

6. Experimental Results

This system presents the study of neural network based clustering algorithm: SOM that includes finding appropriate representation for clustering task creating non-overlapping partitions of input. SOM has a time complexity which is linear in the number of data and input learning rate parameter, but are thought to produce clusters of nearly manual result. If the learning rate is small, the calculation is long and the output result is more accurate. The fuzzy partition of a data set is obtained by FCM algorithm. FCM is an overlapping cluster model, thus it allows the input data to belong to more than one cluster simultaneously according to the varying degrees of object membership[2].

Although the FCM algorithm starts with a random initialization of the partition matrix, it generally converges quite rapidly to accurate and robust in clustering by using the termination tolerance ε value. The resulting partition is different according to the different values of fuzziness parameter m`. So m` is needed to adjust appropriate. Experimental results demonstrate that the better clustering result can be reached by setting of the input parameters of the two algorithms and user can test and analyze the accuracy with various input dataset with variable attributes.

The following figures are the computer simulation results of Self-Organizing Map with 3x3 map size and Fuzzy c-means algorithm with 2x2 map size.

Figure 4. SOM 3x3 map result

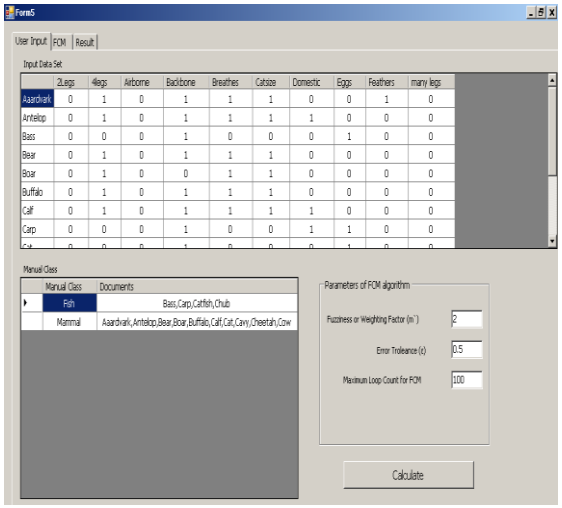

Figure 5. FCM 2x2 map result

The last figure shows the comparison result of SOM and FCM. The system compares the two algorithms based on accuracy.

The accuracy measure for output result is computed as follows:

$$accuracy = \frac{\sum_{i=1}^{n} a_i}{n}$$

n = number of instance in the dataset

a_i = number of instance occurring true positive

In the formula a_i is the actual cluster result of machine learning based results and n is the manual classified data in the dataset. The system is calculated accuracies of two algorithms with the above formula and so user can view with graph.

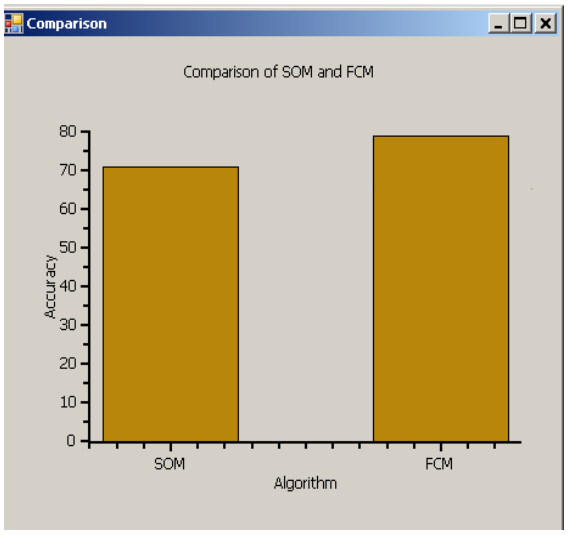

Figure 6. Comparison of FCM and SOM results

7. Conclusion

This system is able to cluster only data that are two dimensions. To calculate accuracy measure, the selected input data must be classified manually. One of the approaches to extend the Fuzzy c-Means (FCM) clustering algorithm is by using the Hyper-spherical Fuzzy c-Means (H-FCM) in which the Euclidean norm is replaced by a dissimilarity function. H-FCM is able to discover good clusters for different levels of fuzziness. Other clustering methods such as K-mean Methods, Bisecting K-mean Methods and frequent term-set association Methods should be used for various datasets to compare clustering methods.

8. References

[1] T.J. Ross, Fuzzy Logic with Engineering Applications, Second Edition, Wiley Student Edition, Wiley India Pvt. Ltd. ISBN 10:81-265-1337-3.

[2] Babuska
"FUZZY CLUSTERING"
URL:http://homes.dsi.unimi.it/~valent/slidecorsi/Bionformatica05
.

[3] J. C. Bezdek
"NUMERICAL TAXONOMY WITH FUZY SETS"
1974, 1: 57-71

[4] J. C. Bezdek, R. Ehrlich, and W. Full
"FCM: FUZZY C-MEANS ALGORITHM"
Computers and Geoscience, 1984.

[5] L. Kaufman and P. J. Rousseeuw
"FINDING GROUPS IN DATA:AN INTRODUCTION TO CLUSTER ANALYSIS"
John Wiley & So, 1990

[6] M.E.S. Mendes and L. Sacks
"DYNAMIC KNOWLEDGE REPRESENTATION FOR TO CLUSTER ANALYSIS"
Sacks Department of Electronic and Electrical Engineering, University College Londaon, Torring Place, London WC1E7JE, UK

[7] **"EMPIRICAL EVALUATION OF CLUSTERING ALGORITHMS"**
Austrian Institute: for East- and Southeast Europe as part of the CLUE Project

[8] R. Mayer
"TEXT MINING WITH ADAPTIVE NEURAL NETWORKS"
Wien University, February, 2004.

[9] R.Callan
"THE ESSENCE OF NEURAL NETWORKS"
Prentice Hall, Europe.

[10] S. Haykin
"NEURAL NETWORKS: A COMPREHENSIVE FOUNDATION"
Second Edition, Prentice Hall International, Inc.

[11] B. Kotsiantis and P.E.Pintelas,
"RECENT ADVANCES IN CLUSTERING"