
*IGI PUBLISHING* **ITB15222**
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.igi-global.com

This paper appears in the publication, **Data Mining and Knowledge Discovery Technologies** edited by D. Taniar © 2008, IGI Global

## Chapter V

# Determination of Optimal Clusters Using a Genetic Algorithm

Tushar, Indian Institute of Technology, Kharagpur, India
Shibendu Shekhar Roy, Indian Institute of Technology, Kharagpur, India
Dilip Kumar Pratihar, Indian Institute of Technology, Kharagpur, India

## Abstract

*Clustering is a potential tool of data mining. A clustering method analyzes the pattern of a data set and groups the data into several clusters based on the similarity among themselves. Clusters may be either crisp or fuzzy in nature. The present chapter deals with clustering of some data sets using the fuzzy c-means (FCM) algorithm and the entropy-based fuzzy clustering (EFC) algorithm. In the FCM algorithm, the nature and quality of clusters depend on the pre-defined number of clusters, level of cluster fuzziness, and a threshold value utilized for obtaining the number of outliers (if any). On the other hand, the quality of clusters obtained by the EFC algorithm is dependent on a constant used to establish the relationship between the distance and*

*similarity of two data points, a threshold value of similarity, and another threshold value used for determining the number of outliers. The clusters should ideally be distinct and at the same time compact in nature. Moreover, the number of outliers should be as minimal as possible. Thus, the previous problem may be posed as an optimization problem, which will be solved using a genetic algorithm (GA). The best set of multi-dimensional clusters will be mapped into 2-D for visualization using a self-organizing map (SOM).*

# Introduction

Clustering is a powerful tool of data mining. Cluster analysis aims to search and analyze the pattern of a data set and group them into several clusters based on their similarity among themselves. It is done in such a way that the data points belonging to a cluster are similar in nature and those belonging to difficult clusters have a high degree of dissimilarity. There exist a number of clustering techniques and those are broadly classified into *hierarchical* and *partitional* methods.

Hierarchical methods iteratively either merge a number of data points into one cluster (called agglomerative method) or distribute the data points into a number of clusters (known as divisive method). An agglomerative method starts with a number of clusters that is equal to the number of data points so that each cluster contains one data point. At each iteration, it merges the two closest clusters into one and ultimately one cluster will be formed consisting of all the data points of them. A divisive method begins with a single cluster containing all the data points. It iteratively divides the data points into more number of clusters and ultimately each cluster will contain only one data point.

The aim of using the partitional methods is to partition a data set into some disjoint subsets of points, such that the points lying in each subset are as similar as possible. Partitional methods of clustering are further sub-divided into hard clustering and fuzzy clustering techniques. In hard clustering, the developed clusters will have their well-defined boundaries. Thus, a particular data point will belong to one and only one cluster. On the other hand, in fuzzy clustering, a particular data point may belong to the different clusters with different membership values. It is obvious that the sum of membership values a data point with various clusters will be equal to 1.0.

This chapter deals with fuzzy clustering. There exist a number of fuzzy clustering algorithms and out of those, the fuzzy c-means (FCM) algorithm (Bezdek, 1981; Dunn, 1974) is the most popular and widely used one due to its simplicity. The performance of the FCM algorithm depends on the number of clusters considered, level of fuzziness and others. However, it has the following disadvantages: