Egyptian Informatics Journal (2016) 17, 183–188

Cairo University

Egyptian Informatics Journal

www.elsevier.com/locate/eij
www.sciencedirect.com

ORIGINAL ARTICLE

Performance comparison of fuzzy and non-fuzzy classification methods

B. Simhachalam [a,*], G. Ganesan [b]

[a] Department of Mathematics, GIT, GITAM University, Visakhapatnam, Andhra Pradesh 530045, India
[b] Department of Mathematics, Adikavi Nannaya University, Rajahmundry, Andhra Pradesh 533296, India

Received 27 July 2015; revised 23 September 2015; accepted 30 October 2015
Available online 29 November 2015

KEYWORDS

k-means;
Fuzzy c-means;
Gustafson–Kessel;
Classification;
Partition based clustering

Abstract In data clustering, partition based clustering algorithms are widely used clustering algorithms. Among various partition algorithms, fuzzy algorithms, Fuzzy c-Means (FCM), Gustafson–Kessel (GK) and non-fuzzy algorithm, k-means (KM) are most popular methods. k-means and Fuzzy c-Means use standard Euclidian distance measure and Gustafson–Kessel uses fuzzy covariance matrix in their distance metrics. In this work, a comparative study of these algorithms with different famous real world data sets, liver disorder and wine from the UCI repository is presented. The performance of the three algorithms is analyzed based on the clustering output criteria. The results were compared with the results obtained from the repository. The results showed that Gustafson–Kessel produces close results to Fuzzy c-Means. Further, the experimental results demonstrate that k-means outperforms the Fuzzy c-Means and Gustafson–Kessel algorithms. Thus the efficiency of k-means is better than that of Fuzzy c-Means and Gustafson–Kessel algorithms.

1. Introduction

Many organizations generate and store large volume of data in their databases. The methods to extract the most useful knowledge from the databases are known as Data mining or knowledge discovery in databases (KDD). Data mining is an analytic process of discovering valid, unsuspected relationships among datasets and transforms the data into a structure that are both understandable and useful to the users.

Data analysis contains several techniques and tools for handling the data. Classification or clustering is well known method in data analysis. It is a multivariate analysis technique to partition the dataset into groups (classes or clusters) in a dataset such that the most indiscernible objects belong to the same group while the discernible objects in different groups. Clustering methods are used as a common technique in many fields such as pattern recognition, machine learning, image segmentation, medical diagnostics and bio-informatics [5].

* Corresponding author. Tel.: +91 9866118074.
E-mail addresses: drbschalam@gmail.com (B. Simhachalam), prof.ganesan@yahoo.com (G. Ganesan).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.eij.2015.10.004
1110-8665

The two important features in clustering are partition-based clustering and hierarchical-based clustering. Partition-based clustering algorithms have the capable of discovering underlying structures of clusters by using appropriate objective function [15]. The algorithms k-means (KM), Fuzzy c-Means (FCM) and Gustafson–Kessel (GK) clustering algorithms are widely used partition-based clustering algorithms. The algorithms k-means and Fuzzy c-Means are proposed based on Euclidean distance measure and an adaptive distance measure was proposed in Gustafson–Kessel (GK) clustering algorithm.

Several comparisons are carried out by the following researchers: Jaindong, Hongzan, Jaiwen, Qiyong [16] analyzed the performance of k-means and Fuzzy c-Means algorithms and reported that the k-means method is preferable to FCM for Arterial Input Function (AIF) detection using both clinical and simulated data. Velmurugun [14] has compared the clustering performance of k-means and Fuzzy c-Means algorithms using different shapes of arbitrary distributed data points and reported that the k-means performs better than FCM. Simhachalam and Ganesan [12] analyzed the performance of Fuzzy c-Means and Gustafson–Kessel algorithms on medical diagnostics systems and reported that the performance of GK method is better than the FCM method. Wang and Garibaldi [17] have compared the performance of k-means and Fuzzy c-Means algorithms on Infrared spectra collected from auxiliary lymph node tissue section. Mousumi Gupta [8] proposed data scaling method in Gustafson–Kessel algorithm for target detection on scaled data and compared with FCM method. Neha and Seema [9] examined the performance between FCM and GK using cluster validity measures. Dibya Joyti and Anil kumar Gupta [3] evaluated the performance between k-means and Fuzzy c-Means algorithms based on time complexity. Soumi Gosh and Sanjay Kumar Dubey [13] evaluated the clustering performance of k-means and Fuzzy c-Means algorithms on the basis of the efficiency of the clustering output and the computational time and reported that k-means is superior to FCM. Bharati and Gohokar [1] compared the color image segmentation performance between k-means and Fuzzy c-Means algorithms.

The work in this paper aimed to compare the performance of the three clustering techniques, k-means (KM), Fuzzy c-Means (FCM) and Gustafson–Kessel (GK). The most popular real world date sets such as Liver Disorders and Wine are applied to test the performance of these algorithms and a comparative analysis is presented in this work. The rest of this work is organized as follows: In Section 2, concise details of data sets and the three algorithms are presented. In Section 3, results and discussion are presented and the conclusions are in Section 4.

2. Materials and methods

Clustering is an unsupervised data analysis which is used to partition a set of records or objects into clusters or classes with similar characteristics. The partition is done in such a fashion that most similar (or related) objects are placed together, while dissimilar (or unrelated) objects are placed in different classes or groups.

The desired characteristics of clustering methods are ability to deal with different types of attributes with high dimensionality, effective handling of outliers and noise with minimum knowledge, ability to discover the underlying shapes and structures of the data, scalability, usability and interpretability. Clustering methods are categorized into five different methods: partitioning method, hierarchical method, data density based method, grid based method and model based or soft computing methods. Among these five methods partition based methods, k-means (KM), Fuzzy c-Means (FCM) and Gustafson–Kessel (GK) clustering algorithms are implemented using two well known data sets liver disorders and wine to generate two clusters and three clusters respectively.

2.1. The dataset

The real world data sets Liver Disorder and Wine were obtained from the UCI Machine Learning Repository donated by Richard [11] and Forina [4] respectively. The Liver data set contains 341 samples with 6 attributes or blood tests each. These blood tests are capable of detecting liver disorders which might arise due to excessive alcohol consumption. The attributes are the measurements of the blood tests namely mean corpuscular volume (mcv), alkaline phosphatase (alkphos), alanine aminotransferase (sgpt), aspartate aminotransferase (sgot), gamma-glutamyl transpeptidase (gammagt) and the number of half-pint equivalents of alcoholic beverages drunk per day (drinks). The 341 samples are clustered into two different classes according to the liver disorders: Class 1 containing 142 samples and Class 2 containing 199 samples. The Wine data set contains 178 samples and each sample has 13 attributes or chemical analysis of the wine derived from three different cultivars but grown in the same region in Italy. The samples are grouped into three different classes according to the cultivars: Cultivar 1 containing 59 samples, Cultivar 2 containing 71 samples and Cultivar 3 containing 48 samples. The attributes are the values of chemical analysis of Alcohol, Malic acid, Ash, Alkalinity of ash, Magnesium, Total phenols, Flavonoids, Nonflavonoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines and Proline.

2.2. k-means clustering

MacQueen [7] introduced the k-means or Hard C-Means algorithm in 1967. It is a partitioning algorithm applied to classify data into $c(1 \leqslant c \leqslant N)$ clusters and each object (observation) can only belong to one cluster at any one time. Consider a dataset Z with N observations. Each observation is an n-dimensional row vector, $z_k = [z_{k1}, z_{k2}, \ldots z_{kn},] \in \Re^n$. The dataset Z is represented as $N \times n$ matrix. The rows of Z represent samples (observations) and the columns are measurements for these samples (objects). k-means model achieves its partitioning by the iterative optimization of its objective function (a squared error function) given as

$$J(V) = \sum_{i=1}^{c}\sum_{k=1}^{N}\|z_k - v_i\|^2 \tag{1}$$

where $\|z_k - v_i\|^2$ is the Euclidean distance calculated between kth object, z_k and ith centroid, v_i. The algorithm comprises the following basic steps: