Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems 253 (2014) 122–137

FUZZY sets and systems

www.elsevier.com/locate/fss

A novel cluster validity index for fuzzy clustering based on bipartite modularity

Dawei Zhang [a], Min Ji [a], Jun Yang [b], Yong Zhang [a], Fuding Xie [b,c,*]

[a] *School of Computer Science and Technology, Liaoning Normal University, Liaoning, Dalian 116081, PR China*
[b] *School of Urban and Environmental Science, Liaoning Normal University, Liaoning, Dalian 116029, PR China*
[c] *Academy of Mathematics and System Sciences, Chinese Academy of Science, Beijing 100080, PR China*

Received 22 November 2012; received in revised form 18 December 2013; accepted 21 December 2013

Available online 28 December 2013

Abstract

A novel cluster validity index whose implementation is based on the membership degrees and improved bipartite modularity of bipartite network is proposed for the validation of partitions produced by the fuzzy c-means (FCM) algorithm. FCM algorithm is employed to group the dataset in order to obtain the membership degree of samples. Then, a weighted bipartite network is constructed by samples and centroids of each cluster. This allows the introduction of a new measurement for optimizing the numbers of clusters for fuzzy partitions. The proposed index utilizes the optimum membership as its global property and the modularity of bipartite network as its local independent property. The proposed index is compared with a number of popular validation indices on fifteen datasets. The experimental results show that the effectiveness and reliability of the proposal is superior to other indices.

Keywords: Fuzzy clustering; Cluster validity index; Bipartite modularity; FCM algorithm

1. Introduction

Clustering has become very important in areas like data mining, pattern recognition, engineering and so on. The purpose of clustering is to divide a given data set into groups (clusters), such that all data in the same group are similar to each other, while data from different clusters are dissimilar. Since clustering is an unsupervised classification process, it has no priori information of data set. A wide variety of clustering algorithms have been proposed in the past decades. Generally speaking, these algorithms can be divided into two classes: hard (crisp) cluster and soft (fuzzy) cluster. Hard clustering algorithms are based on classical set theory and require that a datum either does or does not belong to a cluster. Fuzzy clustering algorithms allow objects to belong to several clusters simultaneously, with different degrees of membership. The results of these clustering algorithms, however, depend on input parameters. For instance, c-means and FCM algorithms require a cluster number c to be predefined. In this case, the question is: What is the optimal cluster number? To answer this question, currently, cluster validity indices research has drawn

[*] Corresponding author at: School of Urban and Environmental Science, Liaoning Normal University, Liaoning, Dalian 116029, PR China.
E-mail address: Xiefd@lnnu.edu.cn (F. Xie).

0165-0114/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.fss.2013.12.013

considerable attention in data mining. Many different cluster validity indices have been defined without any prior class knowledge. In fact, clustering validation is a technique to find a set of clusters that best fits natural partitions without any class information. Up to now, no popular index is known to be suitable for all data sets. Thus, how to define a good validity index to detect a optimal cluster number c is still a challenging problem.

Based on FCM algorithm and the modularity of weighted bipartite network, we introduce a novel cluster validity index to find a proper cluster number. To achieve this goal, the problem of clustering data is converted into detecting the community structures of bipartite network. Afterwards, the modularity of bipartite network is improved. Then we apply the improved modularity successfully to evaluate the quality of the partitions. In next section, we present different cluster validity indices after introducing the fuzzy clustering. Section 3 briefly describes the bipartite network and modularity function. The novel cluster index is proposed in Section 4. Section 5 validates the proposed index. Finally, conclusions and remarks are presented in Section 6.

2. Background

2.1. The fuzzy c-means clustering algorithm

The objective of fuzzy clustering is to partition a data set into c distinct clusters. The well-known fuzzy c-means algorithm proposed by Dunn [1], then extended by Bezdek [2] and its various variations are probably the most commonly used fuzzy clustering methods.

Let $X = \{x_1, \dots, x_n\}$ be a n points data set in a P-dimensional feature space R^P, $X \subset R^P$. The FCM clustering algorithm partitions X into $1 < C < n$ fuzzy groups by minimizing objective function J_m which is the weighted sum of squared errors within groups and is defined as follows:

$$J_m(U, V, X) = \sum_{j=1}^{n} \sum_{i=1}^{C} u_{ij}^m \|x_j - v_i\|_A^2, \quad 1 < m < \infty, \tag{1}$$

and subject to

$$u_{ij} \in [0, 1], \quad \sum_{i=1}^{C} u_{ij} = 1,$$

where $U = [u_{ij}]_{C\times n}$ is a fuzzy partition matrix composed of the membership degree of data point x_j to ith cluster; $V = (v_1, v_2, \dots, v_c)$ is a vector of unknown cluster prototype (centers), $v_i \in R^P$. Norm matrix A defines a measure of similarity between a data point and the cluster prototypes. The parameter m controls the fuzziness of membership of each datum. The cluster centers and the respective membership functions, which are solutions of the constrained optimization problem in Eq. (1), are given by the following equations:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_j - v_i\|_A}{\|x_j - v_k\|_A}\right)^{\frac{2}{m-1}}}, \quad 1 \leqslant i \leqslant C,\ 1 \leqslant j \leqslant n, \tag{2}$$

and

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}, \quad 1 \leqslant i \leqslant C. \tag{3}$$

Eqs. (2) and (3) constitute an iterative optimization procedure.

The FCM algorithm is executed in the following steps:

Step 1: Given a preselected C clustering centers set V and fuzzy factor m ($m > 1$), initialize the fuzzy partition matrix U as Eq. (2).

Step 2: Calculate the fuzzy clustering centroid matrix V by Eq. (3).

Step 3: Use Eq. (2) to update the fuzzy membership matrix U.

Step 4: If the improvement in $J_m(U, V, X)$ is less than a certain threshold (ε), then stop; otherwise go to Step 2.

The FCM algorithm detects clusters that have centroid prototypes of a roughly same size. The Gustafson–Kessel (GK) algorithm is an extension of the FCM, which can detect cluster of different orientations and shapes in a data set