Knowledge-Based Systems 64 (2014) 81–93

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A new cluster validity measure based on general type-2 fuzzy sets: Application in gene expression data clustering

Abolfazl Doostparast Torshizi, Mohammad Hossein Fazel Zarandi *

Department of Industrial Engineering, Amirkabir University of Technology (Tehran Polytechnic), Tehran 15875-5513, Iran

ARTICLE INFO

Article history:
Received 10 November 2013
Received in revised form 28 January 2014
Accepted 25 March 2014
Available online 1 April 2014

Keywords:
General type-2 fuzzy sets
General type-2 fuzzy c-means
Similarity measure
Cluster validity index
Gene expression clustering

ABSTRACT

As a widespread pattern recognition technique, clustering has been widely used in various disciplines including: science, engineering, medicine, etc. One the latest progresses in this field is introduction of general type-2 fuzzy sets and the new clustering method represented on its basis called general type-2 fuzzy c-means. In this paper, the aim is to develop a robust and accurate similarity measure between general type-2 fuzzy sets. Utilizing philosophy behind this developed similarity measure, the first exclusively developed general type-2 fuzzy cluster validity index will be proposed to be used for finding the optimal number of clusters through using general type-2 fuzzy c-means. To verify quality of the proposed approaches, several heavy computations have been conducted on artificial datasets and also real gene expression datasets. Numerical comparisons reveal robustness and quality of the proposed approach compared to several similar approaches in the literature.

1. Introduction

As a high-throughput method, microarray technology has made it possible to study expression levels of thousands of genes, simultaneously. This technology has been widely applied to various fields such as medical diagnosis, biomedicine, bio-machine, characterizing gene functions and understanding molecular processes [1–5,48,49]. Gene expression values obtained from microarray technology are multi-dimensional with large volume. Hence, it is necessary to apply computational analysis methods for extracting the hidden knowledge in this data. One of such methods is clustering.

As an unsupervised pattern recognition technique, clustering partitions the space into several clusters such that the objects in the same cluster are similar based on some criteria while objects in different clusters are tried to have least similarity with the members of other clusters [6,7]. So far, several standard clustering algorithms such as K-means [8], hierarchical clustering [2], Fuzzy C-Means (FCM) [9], Self Organizing Maps (SOM) [10], Simulated Annealing (SA), and Genetic Algorithm-based (GA) clustering algorithms [6,11–14] have been utilized for clustering gene expression data.

Gene expression values obtained from microarray chips are not always complete and they may involve several missing values. Also, according to experimental variations, the expression values may be noisy. In order to eliminate such variations and deal with noisy values, fuzzy logic can be an appropriate tool. FCM is a widespread fuzzy clustering algorithm which has two inputs including: the data to be clustered and the number of clusters. Then, it partitions the data in such a way that the total membership of each pattern to the entire cluster prototypes becomes 1. Meanwhile, FCM and most of its extensions, which are based on type-1 fuzzy sets, suffer from several drawbacks including [15]:

- The scientific basis for choosing the fuzzifier coefficient is not still clear.
- Number of clusters should be assigned a priori.
- For obtaining the optimal partitioning, initial location of cluster centers should be assigned. Most of the times, the FCM converges to local optimal points and different choices of initial cluster centers may lead to different partitioning.
- FCM-based clustering algorithms are highly sensitive to noise and outliers.

Higher order fuzzy sets such as General Type-2 Fuzzy Sets (GT2 FSs) and Interval Type-2 Fuzzy Sets (IT2 FSs) have been proved of being capable of enhancing uncertainty handling abilities of traditional T1 FSs. So they have been extensively applied to T1 clustering algorithms such as FCM. IT2 FCM was proposed by Hwang and Rhee [16]. Their approach used two fuzzifier coefficients and computed two membership values for each pattern called lower

* Corresponding author. Tel.: +98 21 6454 5378.
E-mail address: zarandi@aut.ac.ir (M.H. Fazel Zarandi).

http://dx.doi.org/10.1016/j.knosys.2014.03.023
0950-7051/© 2014 Elsevier B.V. All rights reserved.

and upper membership values. This property significantly improved vagueness handling capabilities of the FCM. On the other hand, newly developed representation techniques such as α-planes made it possible to extend IT2 FCM into its more flexible version GT2 FCM [17]. GT2 FCM relies on the concept of α-planes [18] where each α-plane is an IT2 FS itself. Therefore, mathematical operations of IT2 FSs can be applied to each α-plane.

Since FCM-based clustering techniques cannot automatically estimate the exact number of clusters then there should be a quality criterion to find the optimal number of clusters such that the obtained clusters be accurate and precise as much as possible. These measures are called Cluster Validity Indices (CVIs). To date, many different CVIs have been introduced to the literature including [15,19–26]. Of course, there are some studies on CVIs which can handle IT2 FCM. As an example in [27], Wu and Mendel present a similarity measure for IT2 FCM which utilizes upper and lower membership functions of IT2 FCM to compute the similarity level between two fuzzy sets. Then, in [15], Fazel Zarandi et al. used the proposed approach in [27] in order to compute pairwise similarity value between clusters and also presented a new CVI which was the summation of the computed similarities among the entire clusters. Recently, Ozkan and Turksen [28] have utilized the uncertainty associated with fuzziness level in order to determine the number of clusters in FCM. Unfortunately, no more CVIs can be found which is specifically developed to deal with IT2 FCM. On the other hand, nearly no CVI has been developed for GT2 FCM so far. Hence, one of our scientific goals in this paper is to develop a robust CVI for GT2 FCM for the first time.

1.1. An overview on similarity measures

Similarity measures between fuzzy sets are important issues in fuzzy logic studies. Although several similarity approaches between type-1 FSs have been presented so far but such approaches for higher order FSs like IT2 FSs are scarce. Most of the similarity measures between IT2 FSs are actually extensions of measures originally developed for type-1 fuzzy sets. Some of these measures can be found in [27,37–40].

Similarity measures between GT2 FSs is scarce. One of the first similarity measures between GT2 FSs was presented by Mitchell [41]. But his approach suffers from several drawbacks including: (1) it does not satisfy symmetry, (2) it does not satisfy reflexivity, (3) its results are not robust, i.e., they may change from experiment to experiment. Another GT2 similarity measure has been introduced in [42,43]. In these approaches, the FOU of FSs being compared should be the same which makes the application of these approaches limited to some specific cases. One of the latest GT2 similarity measures is represented in [29]. This similarity measure utilized the concept of α-planes and applied the similarity measure proposed by Wu and Mendel [27] to each α-plane. Suppose $\widetilde{A}$ and $\widetilde{B}$ to be GT2 FSs. Now, the similarity measure between them is defined as follows:

$$S(\widetilde{A},\widetilde{B}) = \bigcup_{\alpha\in[0,1]} \left\{\alpha|[S_L(\widetilde{A},\widetilde{B},\alpha), S_R(\widetilde{A},\widetilde{B},\alpha)]\right\} \tag{1}$$

where

$$S_L(\widetilde{A},\widetilde{B},\alpha)=\min\left\{\frac{\int_{x\in X}\min(S_L^{\widetilde{A}}(x|\alpha),S_L^{\widetilde{B}}(x|\alpha))dx}{\int_{x\in X}\max(S_L^{\widetilde{A}}(x|\alpha),S_L^{\widetilde{B}}(x|\alpha))dx},\frac{\int_{x\in X}\min(S_R^{\widetilde{A}}(x|\alpha),S_R^{\widetilde{B}}(x|\alpha))dx}{\int_{x\in X}\max(S_R^{\widetilde{A}}(x|\alpha),S_R^{\widetilde{B}}(x|\alpha))dx}\right\} \tag{2}$$

$$S_R(\widetilde{A},\widetilde{B},\alpha)=\max\left\{\frac{\int_{x\in X}\min(S_L^{\widetilde{A}}(x|\alpha),S_L^{\widetilde{B}}(x|\alpha))dx}{\int_{x\in X}\max(S_L^{\widetilde{A}}(x|\alpha),S_L^{\widetilde{B}}(x|\alpha))dx},\frac{\int_{x\in X}\min(S_R^{\widetilde{A}}(x|\alpha),S_R^{\widetilde{B}}(x|\alpha))dx}{\int_{x\in X}\max(S_R^{\widetilde{A}}(x|\alpha),S_R^{\widetilde{B}}(x|\alpha))dx}\right\} \tag{3}$$

This approach seems to be strong but there are still some limitations. Suppose $\widetilde{A}$ and $\widetilde{B}$ to be GT2 FSs. If the primary membership functions of both these GT2 FSs are equally increased or decreased then their similarity values should not change. Meanwhile, in this approach the similarity value between two GT2 FSs is based on the location of the primary membership values instead of the difference between them. This means the similarity values between two GT2 FSs varies by transferring them on the plane whose axes are primary domain and primary membership function. In order to make this clear, let us analyze it mathematically.

Consider two GT2 FSs, $\widetilde{A}$ and $\widetilde{B}$ where their primary domains are discretized into N points. Now, consider two α-planes $\widetilde{A}_\alpha$ and $\widetilde{B}_\alpha$. Similarity value between these α-planes can be computed using (2) and (3). If the entire membership values of these α-planes are increased by H, then the lower and upper similarity values between them is computed as below:

$$\begin{aligned} S_L(\widetilde{A},\widetilde{B},\alpha)(S_R(\widetilde{A},\widetilde{B},\alpha)) &= \min(\max)\left\{\begin{array}{c}\dfrac{\sum_{i=1}^{N}\min(S_L^{\widetilde{A_\alpha}}(x_i|\alpha)+H,S_L^{\widetilde{B_\alpha}}(x_i|\alpha)+H)}{\sum_{i=1}^{N}\max(S_L^{\widetilde{A_\alpha}}(x_i|\alpha)+H,S_L^{\widetilde{B_\alpha}}(x_i|\alpha)+H)},\\ \dfrac{\sum_{i=1}^{N}\min(S_R^{\widetilde{A_\alpha}}(x_i|\alpha)+H,S_R^{\widetilde{B_\alpha}}(x_i|\alpha)+H)}{\sum_{i=1}^{N}\max(S_R^{\widetilde{A_\alpha}}(x_i|\alpha)+H,S_R^{\widetilde{B_\alpha}}(x_i|\alpha)+H)}\end{array}\right\} \\ &= \min(\max)\left\{\begin{array}{c}\dfrac{\sum_{i=1}^{N}(\min(S_L^{\widetilde{A_\alpha}}(x_i|\alpha),S_L^{\widetilde{B_\alpha}}(x_i|\alpha)))+NH}{\sum_{i=1}^{N}(\max(S_L^{\widetilde{A_\alpha}}(x_i|\alpha),S_L^{\widetilde{B_\alpha}}(x_i|\alpha)))+NH},\\ \dfrac{\sum_{i=1}^{N}(\min(S_R^{\widetilde{A_\alpha}}(x_i|\alpha),S_R^{\widetilde{B_\alpha}}(x_i|\alpha)))+NH}{\sum_{i=1}^{N}(\max(S_R^{\widetilde{A_\alpha}}(x_i|\alpha),S_R^{\widetilde{B_\alpha}}(x_i|\alpha)))+NH}\end{array}\right\}\end{aligned} \tag{4}$$

It can be observed in (4) that the similarity values between $\widetilde{A}$ and $\widetilde{B}$ will be increased when the primary membership values for the entire primary domains boost up. The same situation occurs when the primary membership values for each α-plane are decreased and as the result, the similarity values between GT2 FSs also decrease which is an unwanted and unreasonable property. Such transferring for two α-planes $\widetilde{A}_\alpha$ and $\widetilde{B}_\alpha$ is illustrated in Fig. 1.

Based on the abovementioned discussions, in this section we present a new distance-based similarity measure between GT2 FSs and prove some of it properties.

1.2. An overview on cluster validity indices

To the best of knowledge of the authors, no CVI specifically designed for GT2 FCM has been proposed so far. Therefore in this section, on the basis of the similarity measure presented in the previous section, we aim to develop a new CVI to find the optimal number of clusters constructed by GT2 FCM.

Most of CVIs in the literature such as [2,19–22,50] concentrate on FCM and its extension. Such CVIs try to satisfy two different clustering concepts [15]:

1. *Compactness:* which represents the similarity between patterns existing in each cluster.
2. *Separability:* Dissimilarity between patterns in different clusters.

A good clustering algorithm produces compact clusters while maximizing the separation between patterns in different clusters. Based on these two clustering concepts, CVIS can be categorized into two groups: *ratio type* and *summation type*. The former type operates based on ratio of compactness to separability while the latter couples these concepts in another way. Another category for finding the optimal number of clusters focuses on the degree of sharing between two clusters. For example, Kim et al. [23] pres-