Pattern Recognition 48 (2015) 933–940

Contents lists available at ScienceDirect

Pattern Recognition

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

Relative entropy collaborative fuzzy clustering method

M. Zarinbal [a,1], M.H. Fazel Zarandi [a,c,*,1], I.B. Turksen [b,c]

[a] Department of Industrial Engineering, Amirkabir University of Technology, 424 Hafez Ave, P.O. Box 15875-4413, Tehran, Iran
[b] TOBB Economics and Technology University, Ankara, Turkey
[c] Knowledge Intelligent Systems Laboratory, University of Toronto, Toronto, Canada

ARTICLE INFO

Article history:
Received 28 January 2014
Received in revised form
14 September 2014
Accepted 17 September 2014
Available online 30 September 2014

Keywords:
Fuzzy set theory
Horizontal and vertical collaborative fuzzy clustering
Relative entropy
Relative entropy collaborative fuzzy clustering

ABSTRACT

The main task of clustering methods, especially fuzzy methods, is to find whether natural grouping exists in data and to impose identity on them. In some situations, data are stored in several data sites and to discover the global structures, clustering methods have to be aware of dependencies in all data sites. Collaborative fuzzy clustering methods have been proposed and widely studied to answer such need. In this paper, a novel collaborative fuzzy clustering method is proposed. In this method, relative entropy concept is used as the communication method, a new approach is applied to calculate the interaction coefficient between data sites, and horizontal and vertical modes of the proposed method are discussed. Performance of the proposed method is evaluated using several experiments and the results show that it has the highest quality of collaboration and could classify data more efficiently.

1. Introduction

Imposing identity on data and finding regularities are the main tasks of pattern recognition techniques, which can be done in supervised or unsupervised manner. In unsupervised or clustering methods, the main goal is to find the natural groupings exist in data [1] and it is assumed that data can only belong to one cluster. In reality, however, data can belong to more than just one cluster with some degree of belonging. This is effectively modeled by fuzzy logic. Many fuzzy clustering methods have been developed ([2–8]) and they have been applied in various areas ([9–13]).

In some situations, such as banking institutions, data have been stored in different data sites with same or different patterns and features. Clearly, having knowledge about the dependencies in all data sites is essential for discovering the global structures [14]. Collaborative fuzzy clustering (CFC) method was proposed to answer such need. CFC generally performs in two phases: fuzzy c-means (FCM) method runs independently at each data site in the first phase and in the second phase, data sites communicate their findings and each data site proceeds with its optimization by focusing on local data [15].

While the concepts of CFC have been widely studied ([16–23]), this paper proposes a number of novel concepts; as the ways of communication affect the results, in this paper a new communication method is proposed using relative entropy (RE) concept. RE enables the method to cluster data and handle noisy datasets more efficiently. That is, as discussed in [1], FCM divides data into given cluster numbers regardless of being noisy or not, whereas it is more natural for noise objects to have very low membership degrees in all clusters. The interaction coefficient between data sites is another important issue that, in most cases, has to be estimated beforehand. In this study, a new approach is applied to calculate this value. Data sites could also have same or different number of data patterns and features. Thus, in this paper, two modes of the proposed method, horizontal and vertical, are discussed and their performances are evaluated using several experiments. The obtained results are then compared with CF [14], CFC [15], CFC-c^*[21], CFC with fixed β, CFC-β_f-c^*, [21], and CFC with dynamic β, CFC-β_d, [21].

The rest of this paper is organized as follows: Section 2 discusses the CFC methods presented in literature. The proposed clustering method and its two modes are discussed in Section 3. Performance of the proposed method and some more discussions are addressed in Section 4. Conclusion is stated in Section 5. Finally, Appendix provides the needed proofs.

2. Related works

The CFC method, first proposed by [16], is a fuzzy clustering method concerns with the extension of fuzzy clustering to several data sites. This is done in two phases: in the first phase, FCM runs

[*] Corresponding author at: Department of Industrial Engineering, Amirkabir University of Technology, Tehran, Iran.

E-mail addresses: mzarinbal@aut.ac.ir (M. Zarinbal), zarandi@aut.ac.ir (M.H. Fazel Zarandi), bturksen@etu.edu.tr (I.B. Turksen).

[1] Tel.: +98 2164541.

http://dx.doi.org/10.1016/j.patcog.2014.09.018
0031-3203/© 2014 Elsevier Ltd. All rights reserved.

independently at each data site and in the second phase, the findings are communicated and each data site proceeds with optimization by focusing on its local data. CFC has two important modes, vertical and horizontal. In horizontal CFC (HCFC), each data site has the same data patterns in different feature spaces, whereas in vertical CFC (VCFC), data sites consist of different data in the same feature space. Thus, data sites communicate the membership degrees and prototypes in HCFC and VCFC, respectively [14].

That is, in HCFC, for clustering N data pattern stored in Pdata sites, optimization of the second phase proceeds with [16]

$$\begin{aligned}\min J_{HCFC_1}[ii] = & \sum_{k=1}^{N}\sum_{i=1}^{c} u_{ik}^2[ii] d_{ik}^2[ii] \\ & + \sum_{\substack{kk=1 \\ kk \neq ii}}^{P} \alpha[ii,jj] \sum_{k=1}^{N}\sum_{i=1}^{c} (u_{ik}[ii] - u_{ik}[jj])^2 d_{ik}^2[ii] \\ & S.T. \begin{cases} \sum_{i=1}^{c} u_{ik}[ii] = 1 & \forall k, ii = 1, \ldots, P \\ 0 < \sum_{k=1}^{N} u_{ik}[ii] < N & \forall i, ii = 1, \ldots, P \\ u_{ik}[ii] \in [0,1] & \forall i, k, ii = 1, \ldots, P \end{cases}\end{aligned} \tag{1}$$

where, for iith data site, $u_{ik}[ii]$ is the membership degree of Kth data, $x_k[ii]$, in ith cluster, $d_{ik}[ii]$ is the distance between $x_k[ii]$ and ith cluster's prototype,$v_i[ii]$, and $c$is the number of clusters. $\alpha[ii,jj]$ is the non-negative entry of collaborative matrix and describes the intensity of interaction between iith and jjth data sites [16].

In VCFC, the data sites consists of different data pattern, $N[1], \ldots, N[P]$, in the same feature space, so, collaboration phase proceeds with [14]

$$\begin{aligned}\min J_{VCFC_1}[ii] = & \sum_{k=1}^{N[ii]}\sum_{i=1}^{c} u_{ik}^2[ii] d_{ik}^2[ii] \\ & + \sum_{\substack{jj=1 \\ jj \neq ii}}^{P} \beta[ii,jj] \sum_{k=1}^{N[ii]}\sum_{i=1}^{c} u_{ik}^2[ii] ||v_i[ii] - v_i[jj]||^2 \\ & S.T. \begin{cases} \sum_{i=1}^{c} u_{ik}[ii] = 1 & \forall k, ii = 1, \ldots, P \\ 0 < \sum_{k=1}^{N[ii]} u_{ik}[ii] < N[ii] & \forall i, ii = 1, \ldots, P \\ u_{ik}[ii] \in [0,1] & \forall i, k, ii = 1, \ldots, P \end{cases}\end{aligned} \tag{2}$$

where, $\beta[ii,jj]$ is the collaboration coefficient of ith and jjth data sites.

In both modes, $\alpha[ii,jj]$ and $\beta[ii,jj]$ have to be determined in advance. To solve this problem in HCFC, Falcon et al. [17,19] proposed an approach to seek the most suitable values of $\alpha[ii,jj]$. In their approach, initial values of $\alpha[ii,jj]$ are determined using rough threshold method and a particle swarm optimization (PSO) driven tuning process is used to optimize two predefined fitness functions.

To solve this problem in VCFC, a new CFC method is proposed by [15], in which the collaboration phase proceeds with:

$$\begin{aligned}\min J_{VCFC_2}[ii] = & \sum_{k=1}^{N[ii]}\sum_{i=1}^{c} u_{ik}^2[ii] d_{ik}^2[ii] \\ & + \beta \sum_{\substack{jj=1 \\ jj \neq ii}}^{P} \sum_{k=1}^{N[ii]}\sum_{i=1}^{c} (u_{ik}[ii] - \tilde{u}_{ik}[ii|jj])^2 d_{ik}^2[ii] \\ & S.T. \begin{cases} \sum_{i=1}^{c} u_{ik}[ii] = 1 & \forall k, ii = 1, \ldots, P \\ 0 < \sum_{k=1}^{N[ii]} u_{ik}[ii] < N[ii] & \forall i, ii = 1, \ldots, P \\ u_{ik}[ii] \in [0,1] & \forall i, k, ii = 1, \ldots, P \end{cases}\end{aligned} \tag{3}$$

where, β is a non-negative collaboration number and $\tilde{u}_{ik}[ii|jj]$ is the induced membership degree computed using [15]

$$\tilde{u}_{ik}[ii|jj] = \frac{1}{\sum_{l=1}^{c} \left(\frac{||x_k[ii] - v_i[jj]||}{||x_k[ii] - v_l[jj]||} \right)^2} \tag{4}$$

Using CFC formulation in Eq. (2), four data-driven approaches is proposed by [21] to improve CFC. For vertical collaboration, CFC-β_f and CFC-β_d are proposed for known c and CFC -$\beta_f - c^*$ for unknown c. CFC-c^* is also proposed for horizontal collaboration with unknown c. In CFC-β_f and CFC $\beta_d - c^*$, $\beta[ii,jj]$ remains fixed for each pair of data sites during the collaboration phase, whereas CFC-β_d and CFC-c^* dynamically adjust $\beta[ii,jj]$ for every pair of data sites at every collaboration stage. In these methods $\beta[ii,jj]$ is estimated by [21]

$$\beta[ii,jj] = \min \left\{ 1, \frac{J_{CFC_2}[ii]}{\sum_{k=1}^{N[ii]} \sum_{i=1}^{c} \tilde{u}_{ik}^2[ii|jj] d_{ik}^2} \right\} \tag{5}$$

Furthermore, a PSO driven CFC method and a learning approach based on self-organizing map (SOM) are proposed by [20] and [22] for both mode of collaboration and for determining optimum sets of $\alpha[ii,jj]$ and $\beta[ii,jj]$.

Table 1 summarizes the abovementioned approaches.

The effectiveness of CFC depends on the way of communication and the communicated findings, which are membership degrees in horizontal mode and prototypes in vertical mode [15]. Hence,

Table 1
Collaborative clustering methods

CFC Mode	Authors	Proposed approach
Horizontal	Falcon et al. [17,19]	Rough threshold method is used to determine $\alpha[ii,jj]$ and a particle swarm optimization driven tuning process is used to optimize the two predefined fitness functions.
	Depaire et al. [20]	A particle swarm optimization driven CFC is proposed to determine $\alpha[ii,jj]$.
	Coletta et al. [21]	CFC-c^* method is proposed for unknown c.
	Ghassany et al. [22]	A learning approach based on self-organizing map is proposed to estimate the collaboration parameter during the collaboration phase.
Vertical	Pedrycz, and Rai [15]	$\beta[ii,jj]$ is replaced with β, a nonnegative collaboration parameter. The collaboration phase proceeds with Eq. (3).
	Depaire et al. [20]	A particle swarm optimization driven CFC is proposed to determine $\beta[ii,jj]$.
	Coletta et al. [21]	CFC-β_f and CFC-β_d methods are proposed for known c and CFC-$\beta_f - c^*$ method is proposed for unknown c.
	Ghassany et al. [22]	A learning approach based on self-organizing map is proposed to estimate the collaboration parameter.