Knowledge-Based Systems 000 (2016) 1–13

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Interval-valued fuzzy set approach to fuzzy co-clustering for data classification

Van Nha Pham[a,b], Long Thanh Ngo[a,*], Witold Pedrycz[c,d,e]

[a] Department of Information Systems, Faculty of Information Technology, Le Quy Don Technical University, 236 Hoang Quoc Viet, Hanoi, Vietnam
[b] MIST Institute of Science and Technology, 17 Hoang Sam, Hanoi, Vietnam
[c] Department of Electrical and Computer Engineering, University of Alberta, Edmonton T6R 2V4 AB, Canada
[d] Department of Electrical and Computer Engineering, Faculty of Engineering, King Abdulaziz University, Jeddah, 21589, Saudi Arabia
[e] Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

ARTICLE INFO

Article history:
Received 18 October 2015
Revised 23 May 2016
Accepted 25 May 2016
Available online xxx

Keywords:
Fuzzy clustering
Fuzzy co-clustering
Interval type-2 fuzzy sets
Interval-valued fuzzy sets
Data classification

ABSTRACT

Data clustering is aimed at discovering a structure in data. The revealed structure is usually represented in terms of prototypes and partition matrices. In some cases, the prototypes are simultaneously formed using data and features by running a co-clustering (bi-clustering) algorithm. Interval valued fuzzy clustering exhibits advantages when handling uncertainty. This study introduces a novel clustering technique by combining fuzzy co-clustering approach and interval-valued fuzzy sets in which two values of the fuzzifier of the fuzzy clustering algorithm are used to form the footprint of uncertainty (FOU). The study demonstrates the performance of the proposed method through a series of experiments completed for various datasets (including color segmentation, multi-spectral image classification, and document categorization). The experiments quantify the quality of results with the aid of validity indices and visual inspection. Some comparative analysis is also covered.

1. Introduction

Data clustering concerns unsupervised learning in which we partition data into groups or clusters on the basis of similarity of their features. The techniques which simultaneously cluster both data and their features are referred to as co-clustering algorithm. The applicability and quality of clusters could be augmented by augmenting co-clustering with the concepts and techniques of fuzzy sets. Subsequently, the resulting techniques can be called fuzzy co-clustering. Fuzzy co-clustering follows the same principle as general co-clustering. The difference is that the boundary between any two clusters is described in terms of membership functions rather than characteristic functions. For example, in document classification, fuzzy co-clustering allows any document and word to belong to more than a single co-cluster. Fuzzy co-clustering is suitable for clustering complex data types as multi-dimensional, multi-features, and of large size. Bezdek et al. [1] introduced fuzzy C-Means (FCM) clustering, which nowadays becomes one of the most commonly considered method of fuzzy clustering. Salehi et al. [35] proposed the synergistic combination model based on Particle Swarm Optimization and fuzzy sets to cope with uncertainty and optimize the seed points and expand the hyper-boxes used in granular computing. Bhoyar et al. [2] also proposed modified FCM for color image segmentation using just noticeable difference histogram. Hu et al. [36] proposed a hierarchical cluster ensemble model based on knowledge granulation to provide a new way to deal with the cluster ensemble problem together with ensemble learning application of the knowledge granulation. Recently, type-2 fuzzy sets have been studied and combined with clustering techniques to enhance the abilities of these methods to capture and quantify the aspect of uncertainty [3–5]. Furthermore interval type-2 fuzzy sets have applied to many problems such as land cover classification [6] and color image segmentation [7]. Yeh et al. [8] used interval type-2 fuzzy to data-based system modeling by combining type-2 fuzzy neural network with a hybrid learning algorithm. Mau et al. [9] proposed an interval Type-2 Fuzzy Subtractive Clustering approach to obstacle detection of robot vision using RGB-D camera. Melin and Castillo [10] presented studies and applications of type-2 fuzzy clustering to segmentation, classification, and pattern recognition. Nguyen et al. [32] proposed a genetic type-2 Fuzzy C-means clustering approach, which is developed and applied to the segmentation and classification of M-FISH images. Ngo et al. [30] exploits local spatial information between the pixel and its neighbors to compute the membership degree by using an interval type-2 fuzzy clustering algorithm. Nguyen et al. [31] proposed

* Corresponding author.
E-mail addresses: famvannha@gmail.com (V.N. Pham), ngotlong@mta.edu.vn (L.T. Ngo), wpedrycz@ualberta.ca (W. Pedrycz).

http://dx.doi.org/10.1016/j.knosys.2016.05.049
0950-7051/

kernel interval-valued Fuzzy C-Means clustering (KIFCM) and multiple kernel interval-valued Fuzzy C-Means clustering (MKIFCM) which were built on a basis of the kernel learning method and the interval valued fuzzy sets with intent to overcome some drawbacks existing in the "conventional" Fuzzy C-Means (FCM) algorithm. Clustering of interval-valued data is considered following various approaches [38,39], in which fuzzy k-Means is proposed using adaptive quadratic distances by optimizing a certain adequacy criterion [38]. Another clustering approach [39] for interval-valued data is based on fuzzy c-ordered medoids using Hubers M-estimators and the Yagers order weighted averaging to make it robust to outliers. Xu et al. [37] also referred to interval-valued fuzzy sets by considering weak transitivity index to quantify the transitivity consistency degree of this relation. Dang et al. [33] proposed methods involving interval type-2 fuzzy sets to realize collaborative clustering. Note that interval valued fuzzy set is the special case of interval type-2 fuzzy set when the interval value J_x is the closed interval.

Most of co-clustering algorithms are used to deal with dyadic data, e.g., the document and word co-occurrence frequencies. Co-clustering technique is an alternative clustering method to cluster some data which exhibit complex data structures, such as web, multi-spectral images, hyper-spectral images and alike... A number of pertinent studies on co-clustering have been reported in [11–14]. The first framework of fuzzy co-clustering was proposed by Honda as fuzzy co-clustering modeling (FCCM) [15] for clustering multi-dimensional, multi-feature data. FCCM is an FCM-type co-clustering model, whose goal is to extract co-clusters of objects and features from co-occurrence matrices. Kummamuru et al. [16] proposed an improved fuzzy co-clustering algorithm for clustering documents and keywords (Fuzzy CoDoK). However, Fuzzy CoDoK is vulnerable to outliers because of the algorithms sole reliance on fuzzy C-means-like nature. Tjhi and Chen [17] proposed an improved fuzzy co-clustering algorithm for clustering documents and words by imposing standard partition-like requirements to generate fuzzy word clusters that capture the natural distribution of words, which may be beneficial for further information retrieval. Tjhi and Chen [18] developed a possibilistic fuzzy co-clustering algorithm (PFCC) for automatic categorization of large collections of document. PFCC integrates a possibilistic document clustering technique and a combined formulation of fuzzy word ranking and partitioning into a fast iterative co-clustering procedure. The key idea behind HFCR [19] is the formulation of the dual-partitioning approach for fuzzy co-clustering, replacing the existing partitioning-ranking approach. HFCR adopts an efficient and practical heuristic method that can be shown to be more robust [17] than the dual-partitioning approach. Kanzawa [20] compared imputation strategies in FNM-based and RFCM-based fuzzy co-clustering. Honda et al. [21] improved the performance of recommenders, which come as a combination of content-based and collaborative filtering approaches in a two-layer graph model. Tjhi et al. [22] proposed fuzzy semi-supervised co-clustering algorithm for categorization of large web documents. In this approach, the clustering process is carried out by incorporating some prior knowledge into the fuzzy co-clustering framework and coming in the form of pairwise constraints provided by users. A new cluster validity measure proposed for finding the optimal number of clusters and verifying quality of the proposed approaches which used fuzzy c-means [24,25] and general type-2 fuzzy c-means [34]. Hanmandlua et al. [23] established a new proposal using fuzzy co-clustering model to segment color image data and assess quality of co-clustering based on some validity indices.

In this study, an interval valued fuzzy co-clustering is proposed by combining advantages of fuzzy co-clustering and interval-valued fuzzy sets. The proposed algorithm, called Interval-Valued Fuzzy Co-Clustering algorithm (IVFCoC), is aimed at solving clustering problems in the presence of complex data. The IVFCoC is an extension of fuzzy co-clustering by using two values of the fuzziness coefficient m_1, m_2 to produce FOU corresponding to the upper and lower values of type-2 fuzzy co-clustering. Thus the membership functions assume interval values. Experiments are conducted on various datasets including those of color segmentation, multi-spectral image classification, and document categorization. We show the quality of the experimental results in terms of performance evaluation criteria, accuracy, stability and efficiency. Some comparative analysis is also delivered.

The paper is organized as follow. Section 2 provides a brief background of fuzzy sets and fuzzy co-clustering algorithm. Section 3 introduces the algorithm IVFCoC. Section 4 covers some experiments. Section 5 includes conclusions and future works.

2. Prerequisites

Here, we offer a concise overview of essentials concerning type-2 fuzzy sets and fuzzy co-clustering.

2.1. Interval-valued fuzzy sets

Definition 1. A type-2 fuzzy set, denoted $\tilde{A}$, is characterized by a type-2 membership function (MF)$\mu_{\tilde{A}}(x,u)$ for every $x \in X$, let J_x be a subset of [0,1]. A type-2 MF $\mu_{\tilde{A}}$ is a function from the set {(x, u)|$x \in X$ and u $\in$ J$_x$} to [0,1], i.e.,

$$\tilde{A} = \{((x,u), \mu_{\tilde{A}}(x,u)) | \forall x \in X, \quad \forall u \in [0,1]\} \tag{1}$$

in which $0 \le \mu_{\tilde{A}}(x,u) \le 1$ and

$$\mathrm{J}_x = \left\{u \in [0,1],\ \mu_{\tilde{A}}(x,u) > 0\right\} \tag{2}$$

At each value of x, we say $x = x'$, the 2-D plane whose axes are u and $\mu_{\tilde{A}}(x',u)$ is called a vertical slice of $\mu_{\tilde{A}}(x,u)$. A secondary membership function is a vertical slice of $\mu_{\tilde{A}}(x,u)$. It comes in the form $\mu_{\tilde{A}}(x = x',u)$ for $x \in X$ and u $\in$ J$_x$, i.e.,

$$\mu_{\tilde{A}}(x = x',u) \equiv \mu_{\tilde{A}}(x') = \int_{u \in J_{x'}} f_{x'}(u)/u, u \in [0,1] \tag{3}$$

$$J_{x'} = \left\{u \in [0,1], \mu_{\tilde{A}}(x',u)\right\} \tag{4}$$

in which $0 \le f_x(u) \le 1$.

Definition 2. Uncertainty in the primary memberships of a type-2 fuzzy set $\tilde{A}$ consists of a bounded region that we call the footprint of uncertainty (FOU). It is the union of all primary memberships, i.e., $FOU(\tilde{A}) = \bigcup_{x \in X} J_x$. An upper MF and a lower MF are two type-1 MFs that are bounds for the FOU of a type-2 fuzzy set $\tilde{A}$. The upper MF is associated with the upper bound of FOU($\tilde{A}$), and is denoted $\overline{\mu_{\tilde{A}}}(x),\ \forall x \in X$. The lower MF is associated with the lower bound of FOU($\tilde{A}$), and is denoted $\underline{\mu_{\tilde{A}}}(x),\ \forall x \in X$, i.e.

$$\underline{\mu_{\tilde{A}}}(x) = sup\left\{u | u \in [0,1], \mu_{\tilde{A}}(x,u) > 0\right\} \tag{5}$$

$$\overline{\mu_{\tilde{A}}}(x) = inf\left\{u | u \in [0,1], \mu_{\tilde{A}}(x,u) > 0\right\} \tag{6}$$

Type-2 fuzzy sets are called interval type-2 fuzzy sets [5] if all $\mu_{\tilde{A}}(x,u) = 1$, i.e., such constructs are defined as follows.

Definition 3. Type-2 fuzzy sets $\tilde{A}$ are called interval type-2 fuzzy sets [5] if all $\mu_{\tilde{A}}(x,u) = 1$.

An interval type-2 fuzzy set $\tilde{A}$ is characterized by an interval membership function $\mu_{\tilde{A}}(x,u) = 1$ where $x \in X$ and $u \in J_x \subseteq [0,1]$, i.e.,

$$\tilde{A} = \{((x,u),1) | \forall x \in X, u \in [0,1]\} \tag{7}$$

Please cite this article as: V.N. Pham et al., Interval-valued fuzzy set approach to fuzzy co-clustering for data classification, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.05.049