Applied Soft Computing 24 (2014) 534–542

Contents lists available at ScienceDirect

Applied Soft Computing

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

A rapid fuzzy rule clustering method based on granular computing

Xianchang Wang[a,b,*], Xiaodong Liu[a], Lishi Zhang[b]

[a] Research Center of Information and Control, Dalian University of Technology, Dalian 116024, PR China
[b] School of Sciences, Dalian Ocean University, Dalian 116023, PR China

ARTICLE INFO

Article history:
Received 14 February 2013
Received in revised form 29 June 2014
Accepted 3 August 2014
Available online 13 August 2014

Keywords:
Fuzzy clustering
Granular computing
Fuzzy rule
Fuzzy number
Sample's description

ABSTRACT

Traditionally, clustering is the task of dividing samples into homogeneous clusters based on their degrees of similarity. As samples are assigned to clusters, users need to manually give descriptions for all clusters. In this paper, a rapid fuzzy rule clustering method based on granular computing is proposed to give descriptions for all clusters. A new and simple unsupervised feature selection method is employed to endow every sample with a suitable description. Exemplar descriptions are selected from sample's descriptions by relative frequency, and data granulation is guided by the selected exemplar fuzzy descriptions. Every cluster is depicted by a single fuzzy rule, which make the clusters understandable for humans. The experimental results show that our proposed model is able to discover fuzzy IF–THEN rules to obtain the potential clusters.

1. Introduction

1.1. Clustering

Clustering is one of the most significant research fields in data mining. It aims at partitioning the data into groups of similar objects (samples, patterns). From a machine learning perspective, what clustering does is to find the hidden patterns of the dataset in an unsupervised way, and the resulting system is usually referred to as a data concept. From a practical perspective, clustering plays an outstanding role in data mining applications such as image segmentation [1], computational biology [2,3], web analysis [4], text mining [5], graph clustering [6], and many others.

1.2. State of the art

Clustering methods can be classified into two types: partitional and hierarchical [7]. The partitional approach produces a single partition of the data points, such as the well-known K-means [8] clustering method. And while the hierarchical approach gives a nested clustering result in the form of a dendrogram (cluster tree), from which different levels of partitions can be obtained such as single-link [9], complete-link [10] and average-link [11].

Recently, Frey and Dueck devised a method called Affinity Propagation (AP) [12], which is an unsupervised clustering algorithm based on message-passing techniques. Zhang et al. presented a clustering method called KAP to generate specified *K* clusters based on AP. Fowlkes et al. proposed a spectral grouping method Normalized Cuts [13] to use the Nyström approximation to extend normalized cut. Zelnik-Manor and Perona proposed a method named Self-Tuning spectral clustering (STSC) [14] in which a "local" scale should be used to compute the affinity between each pair of points. Agarwal and Mustafa presented an extension of the K-means clustering algorithm for projective clustering in arbitrary subspaces (KMPC) [15]. Steinley and Hubert proposed an order-constrained K-means cluster analysis through an auxiliary quadratic assignment optimization heuristic (OCKC) [16]. Two distance based clustering methods al-SL [17] are proposed by Patra et al. for arbitrary shaped clusters. DBCAMM [18] is an approach to merge the sub-clusters by using the local sub-cluster density information.

Relaxing this rigidity (crispness) of the partition has constituted in the past a domain of research in the framework of cluster analysis [19]. Many authors have proposed a fuzzy setting as the appropriate approach to cope with this problem. In fuzzy clustering, the well-known fuzzy c-means (FCM) clustering algorithm, which was first proposed by Dunn [20] and then extended by Bezdek [21], is the best-known and has been extensively used in data clustering and related applications. Some new fuzzy clustering methods are proposed, Honda et al. [22] used the fuzzy principal component

* Corresponding author at: Research Center of Information and Control, Dalian University of Technology, No. 2, Linggong Road, Ganjingzi District, Dalian City, Liaoning Province 116024, PR China. Tel.: +86 41184709381.
E-mail addresses: wxcixll@sohu.com, 79810442@qq.com (X. Wang), xdliuros@hotmail.com (X. Liu), lishizhangcc@163.com (L. Zhang).

http://dx.doi.org/10.1016/j.asoc.2014.08.004
1568-4946/

analysis to obtain the cluster indicator, as well as the responsibility weights of samples for the K-means process in order to develop a robust K-means clustering scheme. Tan et al. proposed an improved FCMBP fuzzy clustering method [23] based on evolutionary programming. A proximity fuzzy framework for clustering relational data [24] is presented by Graves et al. A certain knowledge-guided scheme of fuzzy clustering in which the domain knowledge is represented in the form of viewpoints is introduced [25] by Pedrycz et al., the users point of view at data, which are represented in a plain numeric format or through some information granules, is included in the clustering process.

1.3. Contribution and paper organization

Nevertheless, samples are assigned to clusters by these traditionally clustering techniques, users need to manually give descriptions for all clusters. In addition, users sometimes have no specific idea regarding how to explain the clustering results, thus, they might give inappropriate descriptions. A clustering technique is proposed in this study to discover fuzzy IF–THEN rules and provide description of clusters. Each cluster is depicted by a single fuzzy IF–THEN rule. A fuzzy rule-based clustering system is a special case of fuzzy modeling, the acquired knowledge with these system may be more human understandable [26]. The proposed clustering method named FRCGC is based on granular computing. The obtained clusters are specified with some interpretable fuzzy rules, which make the clusters understandable for humans. The experimental results show that our proposed model is able to discover fuzzy IF–THEN rules to obtain the potential clusters.

The remainder of this paper is organized as follows. The following section, we provide an overview of data representation and preliminary notions. The procedure of our clustering algorithm is presented in section "Our proposed rapid fuzzy rule clustering method". Section "Illustrative experiments" provides the detailed analysis of the experiments on a synthetic data. In section "Experimental evaluation", we thoroughly evaluate the efficacy of the proposed model through a number of experiments, using publicly available data. Finally, in the concluding section, we summarize and discuss our results.

2. Preliminary notions

This section briefly reviews some background concepts regarding data representations as well as notions. Let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of n samples, where $x_i \in R^d$ $(i = 1, 2, \ldots, n)$ denotes the ith sample. f_j $(j = 1, 2, \ldots, d)$ denotes the jth column (feature) of X. Thus, $X = (x_{ij})$ is an $n \times d$ matrix representing data, each column of X corresponds to a feature, whereas each row corresponds to a sample.

2.1. Fuzzy number

The non-symmetric trapezoidal and triangular forms of fuzzy numbers [27] are used (Fig. 1) in this paper to construct fuzzy rules. In Fig. 1, feature value is divided into four fuzzy subspaces, "big", "medium big", "medium small", and "small". The fuzzy numbers: trapezoidal for the linguistic terms "big", "small", and triangular for the linguistic term "medium big", "medium small" are used. It is often possible, on the basis of expert experience, to define the parameters of fuzzy number ϕ by means of linguistic variables.

The parameters of fuzzy numbers are defined by the method of Equal Interval Width/Uniform Binning (EIB) [28]. This method relies on sorting the jth feature value and dividing the f_j values into equally spaced bin ranges. A seed K (the number of cluster) supplied by the user determines how many bins are required. With this seed K, it is just a matter of finding the maximum and minimum values of

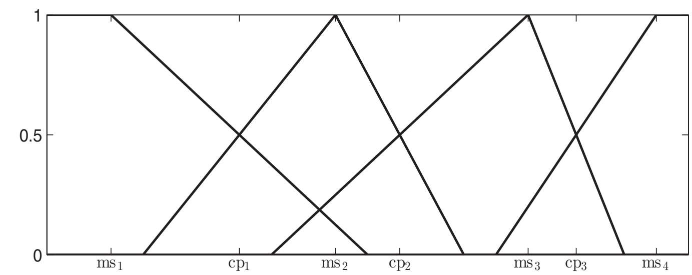

Fig. 1. Triangular and trapezoidal forms of fuzzy number.

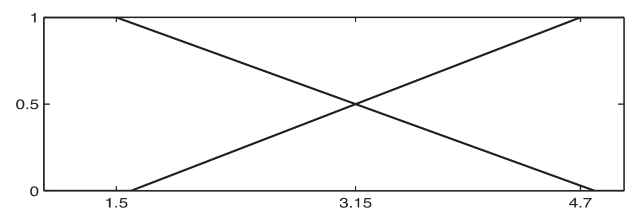

Fig. 2. Fuzzy numbers generated on the feature f_1.

f_j to derive the range and then partition the data into K bins. The bin width is computed by: $\varepsilon = (max(f_j) - min(f_j))/K$, and bin thresholds (cut points) are constructed at $cp_i = min(f_j) + i\varepsilon$, where $i = 1, 2, \ldots, K - 1$. Let ms_k denote the mean of all of the samples that fall into the kth bin.

Example 1. To illustrate the process of generating fuzzy numbers, let us consider the dataset that is shown in Table 1. Assume that $K = 2$, for feature f_1, $min(f_1) = 1.4$, $max(f_1) = 4.9$, $\varepsilon = (max(f_j) - min(f_j))/K = (4.9 - 1.4)/2 = 1.75$, $cp_1 = min(f_j) + \varepsilon = 1.4 + 1.75 = 3.15$, $ms_1 = mean(1.4, 1.6, 1.5, 1.4, 1.5, 1.6) = 1.5$, $ms_2 = mean(4.6, 4.7, 4.9, 4.6) = 4.7$, and the fuzzy numbers are shown in Fig. 2 respectively corresponding to "big" and "small". The membership value of sample belonging to fuzzy number is shown in Table 2, where, ϕ_{11} stands for the value of feature f_1 is "small", ϕ_{21} stands for the value of feature f_1 is "big",ϕ_{12} stands for the value of feature f_2 is "small", ϕ_{22} stands for the value of feature f_2 is "big".

2.2. The description of sample

IF–THEN clustering rules are intuitively comprehensible for most humans since they represent knowledge at a high level of abstraction involving logical conditions rather than point-based cluster representations. In this paper, a clustering rule R defined in the continuous space R^d is knowledge representation in the form:

$$R: \text{ IF } x_i \text{ is } \phi_{11} \text{ and}, \ldots, \text{ and } x_i \text{ is } \phi_{dd} \text{ THEN cluster\_label}$$

Table 1
A small dataset with 10 samples.

Feature	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
f_1	1.4	4.6	1.6	1.5	1.4	4.7	1.5	4.9	1.6	4.6
f_2	0.3	0.2	1.2	0.4	0.5	1.4	0.3	1.5	1.3	1

Table 2
The membership value of samples belonging to fuzzy numbers.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
$\phi_{11}(\cdot)$	1	0.06	0.97	1	1	0.03	1	0	0.97	0.06
$\phi_{21}(\cdot)$	0	0.97	0	0	0	1	0	1	0	0.97
$\phi_{12}(\cdot)$	1	1	0.16	0.94	0.84	0	1	0	0.06	0.35
$\phi_{22}(\cdot)$	0	0	0.91	0	0.09	1	0	1	1	0.67