Knowledge-Based Systems 91 (2016) 189–203

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A tree-based incremental overlapping clustering method using the three-way decision theory

Hong Yu *, Cong Zhang, Guoyin Wang

Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

ARTICLE INFO

Article history:
Received 18 January 2015
Revised 24 April 2015
Accepted 31 May 2015
Available online 6 June 2015

Keywords:
Incremental clustering
Overlapping clustering
Search tree
Three-way decision theory

ABSTRACT

Existing clustering approaches are usually restricted to crisp clustering, where objects just belong to one cluster; meanwhile there are some applications where objects could belong to more than one cluster. In addition, existing clustering approaches usually analyze static datasets in which objects are kept unchanged after being processed; however many practical datasets are dynamically modified which means some previously learned patterns have to be updated accordingly. In this paper, we propose a new tree-based incremental overlapping clustering method using the three-way decision theory. The tree is constructed from representative points introduced by this paper, which can enhance the relevance of the search result. The overlapping cluster is represented by the three-way decision with interval sets, and the three-way decision strategies are designed to updating the clustering when the data increases. Furthermore, the proposed method can determine the number of clusters during the processing. The experimental results show that it can identifies clusters of arbitrary shapes and does not sacrifice the computing time, and more results of comparison experiments show that the performance of proposed method is better than the compared algorithms in most of cases.

1. Introduction

Most of existing clustering algorithms usually analyze static datasets in which objects are kept unchanged after being processed [1,2]. However, in many practical applications, the datasets are dynamically modified which means some previously learned patterns have to be updated accordingly [3,4]. Although these approaches have been successfully applied, there are some situations in which a richer model is needed for representing a cluster [5,6]. For example, a researcher may collaborate with other researchers in different fields, therefore, if we cluster the researchers according to their interested areas, it could be expected that some researchers belong to more than one cluster. In these areas, overlapping clustering is useful and important as well as incremental clustering.

For this reason, the problem of incremental overlapping clustering is addressed in this paper. The main contribution of this work is an incremental overlapping clustering detection method, called TIOC-TWD (Tree-based Incremental Overlapping Clustering method using the Three-Way Decision theory). The proposed method introduces a new incremental clustering framework with three-way decision using interval sets and a new searching tree based on representative points, which together allows to obtain overlapping clusters when data increases. Besides, the TIOC-TWD introduces new three-way strategies to update efficiently the clustering after multiple objects increases. Furthermore, the proposed method can dynamically determine the number of clusters, and it does not need to define the number of cluster in advance. The above characteristics make the TIOC-TWD appropriate for handling overlapping clustering in applications where the data is increasing.

The experimental results show that the proposed method not only can identify clusters of arbitrary shapes, but also can merge small clusters into the big one when the data changes; the proposed method can detect new clusters which might be the result of splitting or new patterns. Besides, more experimental results show that the performance of proposed method is better than the compared algorithms in most of cases. We note that a short version of this work had been appeared in the RSCTC-2014 Workshop on the Three-Way Decisions and Probabilistic Rough Sets [7].

2. Related work

Nowadays, there are some achievements on the incremental clustering approaches. Ester et al. [8] put forward the IncDBSCAN clustering algorithm based on the DBSCAN. After that, Goyal et al. [9]

* Corresponding author. Tel.: +86 13617676007.
E-mail addresses: yuhong@cqupt.edu.cn (H. Yu), zhangcong0214@163.com (C. Zhang), wanggy@cqupt.edu.cn (G. Wang).

http://dx.doi.org/10.1016/j.knosys.2015.05.028
0950-7051/

proposed the derivation work which is more efficient than the IncDBSCAN because it is capable of adding points in bulk to existing set of clusters. Patra et al. [10] proposed an incremental clustering algorithm based on distance and leaders, but the algorithm needs to search the whole data space to find the surrounding leaders. Ibrahim et al. [11] proposed an incremental clustering algorithm which can maximize the relatedness of distances between patterns of the same cluster. Ning et al. [12] proposed an incremental spectral clustering approach by efficiently updating the eigen-system, but it could not find the overlapping clusters. Pensa et al. [13] proposed an incremental hierarchical co-clustering approach, which computes a partition of objects and a partition of features simultaneously but it cannot find out the overlapping clusters.

Meanwhile, some approaches, addressing the problem of incremental overlapping clustering, have been reported. Hammouda and Kamel [14] proposed similarity histogram-based clustering method based on the concept of Histogram Ratio of a cluster. Gil-García and Pons-Porrata [15] proposed dynamic hierarchical compact method and dynamic hierarchical star method, these methods is time consuming due to the framework of hierarchical clustering. Pérez-Suárea et al. [16] proposed an algorithm based on density and compactness for dynamic overlapping clustering, but it builds a large number of small clusters. Lughofer [17] proposed dynamic split-and merge operations for evolving cluster models, which are learned incrementally but can only deal with crisp clustering. Labroche [18] proposed online fuzzy medoid based clustering algorithms, which are adapted to overlapping clusters but the number of clusters need to define in advance.

Therefore, the main objective of this paper is to propose an approach that combine both processing of incremental data and obtaining of overlapping clusters. For this kind of problem, some scholars had pointed out that the clustering approaches to combine with rough sets are impactful [19]. Thus, Parmar et al. [20] proposed an algorithm for clustering categorical data using rough set theory; Chen et al. [21,22] researched the incremental data mining with rough set theory; Peters et al. [23] proposed the dynamic rough clustering; and Lingras et al. [24] reviewed fuzzy and rough approaches for soft clustering.

Further, the three-way decision with interval sets provides an ideal mechanism to represent overlapping clustering. The concept of three-way decisions was developed with researching the rough set theory [25]. A theory of three-way decision is constructed based on the notions of acceptance, rejection and noncommitment, and it is an extension of the commonly used binary-decision model with an added third option [26]. Three-way decision approaches have been successfully applied in decision systems [27–29], email spam filtering [30], three-way investment decisions [31,32], clustering analysis [33], and a number of other applications [25,34]. In our previous work [33], we had proposed a three-way decision strategy for overlapping clustering, where a cluster is described by an interval set. In fact, Lingras and Yan [35] had introduced the concept of interval sets to represent clusters, and Lingras and West [36] proposed an interval set clustering method with rough k-means for mining clusters of web visitors. Yao et al. [37] represented each cluster by an interval set instead of a single set as the representation of a cluster. Chen and Miao [38] described a clustering method by incorporating interval sets in the rough k-means.

In this paper, we propose the three-way decision clustering, which is applicable to crisp clustering as well as overlapping clustering. There are three relationships between an object and a cluster: (1) the object certainly belongs to the cluster, (2) the object certainly does not belong to the cluster, and (3) the object might or might not belong to the cluster. It is a typical three-way decision processing to decide the relationship between an object and a cluster. Objects in the lower bound are definitely part of the cluster, and only belong to that cluster; while objects between the two bounds are possibly part of that cluster and potentially belong to some other clusters.

Furthermore, in the field of incremental learning, it is common to learn from new incremental samples based on the existing results. The tree structures are particularly well suited for this task because they enable a simple and effective way to search and update. At the same time, trees are easy to store the learned patterns (results), which can save lots of duplicate learning time. Tree structures have been successfully used in some typical incremental learning approaches [39,40]. Therefore, this paper will use a tree to store the searching space, where a node of tree indicates the information corresponding to some representative points.

3. Description of the problem

3.1. Three-way decision clustering

To define our framework, let a universe be $U = \{\mathbf{x}_1, \ldots, \mathbf{x}_n, \ldots, \mathbf{x}_N\}$, and the resulting clustering scheme $\mathbf{C} = \{C_1, \ldots, C_k, \ldots, C_K\}$ is a family of clusters of the universe. The $\mathbf{x}_n$ is an object which has D attributes, namely, $\mathbf{x}_n = (x_n^1, \ldots, x_n^d, \ldots, x_n^D)$. The x_n^d denotes the value of the d-th attribute of the object $\mathbf{x}_n$, where $n \in \{1, \ldots, N\}$, and $d \in \{1, \ldots, D\}$.

We can look at the cluster analysis problem from a decision making perspective. For crisp clustering, it is a typical two-way decision; meanwhile for overlapping clustering or soft clustering, it is a type of three-way decision. Let's review some basic concepts of clustering using interval sets from our previous work [33]. In contrast to the general crisp representation of a cluster, where a cluster is a set of objects, we represent a cluster as an interval set. That is,

$$C_k = [\underline{C_k}, \overline{C_k}], \tag{1}$$

where $\underline{C_k}$ is the lower bound of the cluster C_k, $\overline{C_k}$ is the upper bound of the cluster C_k, and $\underline{C_k} \subseteq \overline{C_k}$.

Therefore, we can define a cluster by the following properties:

$$(i)\ \underline{C_k} \neq \varnothing, 0 < k \leq K; \quad (ii)\ \bigcup \overline{C_k} = U. \tag{2}$$

Property (i) implies that a cluster cannot be empty. This makes sure that a cluster is physically meaningful. Property (ii) states that any object of U must belong to the upper bound of a cluster, which ensures that every object is properly clustered.

With respect to the family of clusters, $\mathbf{C}$, we have the following family of clusters formulated by interval sets as:

$$\mathbf{C} = \{[\underline{C_1}, \overline{C_1}], \ldots, [\underline{C_k}, \overline{C_k}], \ldots, [\underline{C_K}, \overline{C_K}]\}. \tag{3}$$

Therefore, the sets $\underline{C_k}, \overline{C_k} - \underline{C_k}$ and $U - \overline{C_k}$ formed by certain decision rules constitute the three regions of the cluster C_k as the positive region, boundary region and negative region, respectively. The three-way decisions are given as:

$$\begin{aligned} POS(C_k) &= \underline{C_k}, \\ BND(C_k) &= \overline{C_k} - \underline{C_k}, \\ NEG(C_k) &= U - \overline{C_k}. \end{aligned} \tag{4}$$

Objects in $POS(C_k)$ definitely belong to the cluster C_k, objects in $NEG(C_k)$ definitely do not belong to the cluster C_k, and objects in the region $BND(C_k)$ might or might not belong to the cluster.

Any data mining technique needs to have a clear and precise evaluation measure. In clustering, evaluations such as the similarity between objects and compactness of clusters are appropriate indicators