New Hierarchical Clustering Method in Interval Data

ABBAS Moncef[1], KERBOUI Roumeissa[2]

[1,2]USTHB, Faculty of Mathematics, Labo. AMCD & RO, algiers
[1] *moncef_abbas@yahoo.com* [2]romahisam@gmail.com

Abstract

In computer science, various clustering algorithms have been developed since conventional hierarchical cluster analysis (HCA), which aims to discover a consistent hierarchy of clusters with different granularities (partitions). In hierarchical clustering, has properties to the representation as tree named dendrogram where each node is associated with merging of two (or more) partitions, where are useful for the interpretation of clustering results and their visualization. Many clustering algorithms have been developed since conventional HCA but are limited in the application. However, the conventional clustering algorithms (mostly HCA) still play an important role. In this paper, we introduce a new ranking based hierarchical clustering on an interval data set with the help of dominance relation, which is efficient to improve ranking clustering accuracy. In each iteration, first a directional distance is calculated on the original data interval. An individu is then created on the selected samples. The quality of clustering is enhaced with small extra computation cost incurred. The extensive experiments indicate that our proposal is superior to distance based hierarchical clustering to obtain in the same time the ranking of clustering. In comparison vaiation measure criterion is also adapted for judging whether the closest pair of clusters can be united or a previous cluster can be split.

Mots-clef: Decision Theory and Analysis, Emerging Applications of OR, Multi-Criteria Decision Analysis.

The contact person (who attend the conference): KERBOUI Roumeissa

1 Introduction

Clustering is a fondamental data analysis tash that creating clusters of similar individus from a given input set. Clustering allows us to identify interesting structures and relation ships within a data set, thus enabling us to make discover or predictions hypotheses to account in the data for the detected structure. Cluster analysis is based to distinct data groups consisting of similar observation. The need in such clustering approach has been seen in various areas of science. In the litterature, many clustering methods proposed can be divided into two main groups: hierarchical methods with different splitting rules called linkages and partition-optimization algorithms such as k-means, Hierarchical clustering [2]. Hierarchical and nonhierrchical clustering algorithms are the two main approachs of clustering algorithms. The number of clusters k is often assumed to be fixed in advance. The inputs to partitional algorithms are the data, a distance metric, and the number of clusters k. The result of each algorithms is a cluster set data which the same individu belong to the same cluster.

One of the most popular algorithm from the family of partitional clustering algorithms is k-means [3]. In many statistical analysis systems focus on distance-based cluster analysis can be used k-means to obtain clusters. But, the disadvantage of users in this method that how fixe the number of clusters k. For this reason,

X	a_1	a_2	...	a_m
x_1	$[\underline{a}_{11}, \overline{a}_{11}]$	$[\underline{a}_{12}, \overline{a}_{12}]$	...	$[\underline{a}_{1m}, \overline{a}_{1m}]$
x_2	$[\underline{a}_{21}, \overline{a}_{21}]$	$[\underline{a}_{22}, \overline{a}_{22}]$	...	$[\underline{a}_{2m}, \overline{a}_{2m}]$
...	...	...		
x_n	$[\underline{a}_{n1}, \overline{a}_{n1}]$	$[\underline{a}_{n2}, \overline{a}_{n2}]$	...	$[\underline{a}_{nm}, \overline{a}_{nm}]$

Table 1: Interval Data

there is really no good solution to predict the optimal number of clusters which is one of the most challenging issues involved with it.

The hierarchical clustering approach can be solve this problem [1]. This approach starts with each input as an individual cluster. The process of clustering will proceed by joining two or more clusters to obtain a new one, until the satisfying of certain termination conditions. The relationship between the input and output alternatives of the hierarchical clustering algorithm, is well represented by a tree, known as *dendrogram*. This representation offer better distribution indifferent abstraction levels, in consequence, the hierarchcal clustering algorithms is an optimale choice for data visualization and exploration.

Our proposed framework, we try to fully applied to incorporate into the existing clusteing algorithm on interval data. With specified metric, the original algorithm merges the items on individual data. But in interval data, our framework enhances the performance by adaptively choosing the directional distance with respect to the next meaningful clustering level. A real data sets proposed in [4], conduct us extensive empirical studies on both synthetic to evaluate the practical values of our proposal. In both hierarchical and data interval clustering, the experimental results show dramatically outperforms of our framework the existing methods on clustering quality.

In this study, we first introduce the concept of directional distance. Then, calculate the directional distance for each alternative to define an ordered mutual information. Based on this consideration, applying to the hierarchical clustering, we propose a principale approach to clustering sets of alternatives with interval values. Finally, we also take a real case proposed in [4] for verifying the effectivity of the proposed approach in this paper.

The paper is organized as follows; Section2: some preliminary concepts and important properties if interval ordered information systems. Section3: introduce the concepts and notion of directional distance index with interval values. Main algorithm of clustering *Hierarchical clustering* has been introduced in section4. In section5, we describe the approach using the combination between hierarchy clustering and Directional Distance approach proposed in this paper.

2 Interval Data

There are several problems where the informations collected for n alternatives $X = \{x_1, x_2, ..., x_n\}$ and each alternative x_i is represented by n interval-type variables $A == \{a_1, a_2, ..., a_n\}$ for m criteria, such that each (x_i) is an interval $[\underline{a}_{ij}, \overline{a}_{ij}]$ and the $i-th$ row $([\underline{a}_{i1}, \overline{a}_{i1}], [\underline{a}_{i2}, \overline{a}_{i2}], ..., [\underline{a}_{im}, \overline{a}_{im}])$ describes the individual x_i (see Table 1)

This is equivalent to considering the m-dimensional interval $x_i = [\underline{a}_{i1}, \overline{a}_{i1}] \times [\underline{a}_{i2}, \overline{a}_{i2}] \times ... \times [\underline{a}_{im}, \overline{a}_{im}]$ in the space $\mathbb{R}^m$. Note that if $\underline{a}_{ij} = \overline{a}_{ij}$ for some (x_i), this means that the underlying cell is single-valued.

3 Preliminaries Concepts

In this section, we review some basic concepts brievly. First, the directional distance index to measure the preferability degree of the object x_i over the object x_j.

Definition 3.1. [4]. Given two interval numbers $f(x_i, a) = [\underline{a}(x_i), \overline{a}(x_i)]$ and $f(x_j, a) = [\underline{a}(x_j), \overline{a}(x_j)]$. *Directional distance index* between two alternatives under the attribute a is defined as

$$DDI_a(x_i, x_j) = \frac{1}{2} + \frac{1}{4}\frac{\overline{a}(x_i) - \overline{a}(x_j) + \underline{a}(x_i) - \underline{a}(x_j)}{max(\overline{a}(x)) - min(\underline{a}(x))}$$

where $max(\overline{a}(x)) = max\{\overline{a}(x_1), \overline{a}(x_2), ..., \overline{a}(x_{|U|})\}$, $min(\underline{a}(x)) = min\{\underline{a}(x_1), \underline{a}(x_2), ..., \underline{a}(x_{|U|})\}$.

In particular, $DDI_a(x_i, x_j) = \frac{1}{2}$ if $max(\overline{a}(x)) = min(\underline{a}(x))$.

In fact, we can make a comparison between two objects. Futhermore, it is no doubt that two objects can be compared under all considered attributes according to the following formula

$$DDI_A(x_i, x_j) = \frac{1}{|A|}\sum_{\forall a \in A} DDI_a(x_i, x_j), \tag{3.1}$$

where $A \subseteq AT$, and $\mid A \mid$ denotes the cardinality of a considered attribute set.

Definition 3.2. [4] Entire directional distance index of each object is defined as

$$DDI_A(x_i) = \frac{1}{\mid U \mid -1}\sum_{i \neq j} DDI_A(x_i, x_j), x_i, x_j \in U. \tag{3.2}$$

Definition 3.3. Let be a matrix of interval data U under a set of criteria $A \subset AT$, The prototypes of a set of alternatives $\overline{y} = [\overline{\alpha}; \overline{\beta}]$ as:

$$\overline{\alpha} = \frac{\max_{i \in A} \underline{a}_{ij} + \min_{i \in A} \underline{a}_{ij}}{2} \; ; \; \overline{\beta} = \frac{\max_{i \in A} \overline{a}_{ij} + \min_{i \in A} \overline{a}_{ij}}{2} \tag{3.3}$$

4 Hierarchical clustering approach

In this paper, a novel conceptual clustering method is presented for created and characterizing each cluster.

4.1 Similarity measure

Given two clusters (C_i, C_j) and their interval feature (F_i, F_j), the distance is typically computed using the minimum different between their directional distance index DDI_A as follows:

A core need is to measure the similarity between two clusters and merged the closest pair of clusters into one larger cluster iteratively. So, merging clusters is calculated as follows:

$$Cluster(C_i; C_j) \Leftrightarrow \min_j [\mid DDI_A(F_i) - DDI_A(F_j) \mid; i \neq j] \tag{4.1}$$

where C_i and C_j are clusters can be merged, $Diff(F_i, F_j)$ is the directional distance index different between C_i and C_j, F_i and F_j are the interval feature of C_i and C_j respectively.

5 Main Work

The main approach is descriped as follows:

Step 1: For each alternative , we calculate their directional distance index descriped in 3.1.

Step 2: Class in the same group, each paire of alternative who are the nearest.

Step3 : For each new cluster, calculate the prototype $c^* = [\alpha^*; \beta^*]$, where $\alpha^*; \beta^*$ defined as 3.3.

Repeat *Step 2 & 3* until obtaining final clustering juste the main cluster, conteint all the set of alternatives.

6 Conclusions

We showed new algorithm performing hierarchy clustering in interval data based to directional distance index. In addition to almost all of other results on hierarchy clustering, the goal is to partition the set of interval data alternative into clusters. Indeed ,dynamic cluster methods for interval data are represented clustering method is considered. The method furnish a partition of the input data and a corresponding prototype for each class by finding each paire alternative of interval closest on the directional distance index for each step. In both methods, the prototype of each class is represented by a vector of interval , where the bounds of these intervals for a variable are, respectively, the average of the set of lower bounds and the average of the set of upper bounds of the intervals of the objects belonging to the class for the same variable. The convergence of these algorithms and the decrease of their partitioning criteria at each iteration is due to the optimization of their adequacy critrion. Experimental result show the advantage of our approach over clustering methods. The algorithm can be viewed as a general tool to clustering interval data problem that globally resume in the Dendrogram.

References

[1] Cai, R., Zhang, Z., Tung, A. K. H., Dai, C., and Hao, Z. A general framework of hierarchical clustering and its applications. *Inf. Sci. 272* (2014), 29–48.

[2] Hong, X., Wang, J., and Qi, G. Comparison of spectral clustering, k-clustering and hierarchical clustering on e-nose datasets: Application to the recognition of material freshness, adulteration levels and pretreatment approaches for tomato juices. *Chemometrics and Intelligent Laboratory Systems 133* (2014), 17 – 24.

[3] Rashedi, E., Mirzaei, A., and Rahmati, M. An information theoretic approach to hierarchical clustering combination. *Neurocomputing 148* (2015), 487–497.

[4] Song, P., Liang, J., and Qian, Y. A two-grade approach to ranking interval data. *Knowl.-Based Syst. 27* (2012), 234–244.