Contents lists available at ScienceDirect

Discrete Applied Mathematics

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

Multilevel clustering models and interval convexities

Patrice Bertrand [a,*], Jean Diatta [b]

[a] *Université Paris-Dauphine, PSL Research Universit, Ceremade, 75775 Paris Cedex 16, France*

[b] *LIM-EA2525, Université de La Réunion, Saint-Denis, France*

ARTICLE INFO

Article history:
Received 4 November 2014
Received in revised form 16 August 2015
Accepted 21 December 2016
Available online xxxx

Keywords:
Abstract convexity
Interval operator
Pyramidal clustering model
k-weak hierarchy
Paired hierarchy

ABSTRACT

The *k*-weakly hierarchical, pyramidal and paired hierarchical models are alternative multilevel clustering models that extend hierarchical clustering. In this paper, we study these various multilevel clustering models in the framework of general convexity. We prove a characterization of the paired hierarchical model via a four-point condition on the segment operator, and examine the case of *k*-weakly hierarchical models for $k \geq 3$. We also prove sufficient conditions for an interval convexity to be either hierarchical, paired hierarchical, pyramidal, weakly hierarchical or *k*-weakly hierarchical. Moreover, we propose a general algorithm for computing the interval convexity induced by any given interval operator, and deduce a unified clustering scheme for capturing either of the considered multilevel clustering models. We illustrate our results with two interval operators that can be defined from any dissimilarity index and propose a parameterized definition of an adaptive interval operator for cluster analysis.

1. Introduction

Cluster analysis aims at finding a collection of subsets of a given ground set, called *clusters*, such that the elements of each cluster are as much as possible closer w.r.t. some dissimilarity function, whereas elements of any two distinct clusters are as much as possible the most dissimilar. Such a collection is accordingly called a *clustering*. There are three main types of clusterings: partitions, coverings and, the most general ones, which we hereafter refer to as multilevel clusterings. In a partition (resp. covering), two clusters cannot (resp. can) overlap but are never nested. In a *multilevel clustering*, two clusters can be nested as well as overlapped or disjoint. Some particular multilevel clustering models are specifically studied in the cluster analysis literature. From the more specific to the more general, the most known of these multilevel clustering models are the *hierarchical, paired-hierarchical, pyramidal and k-weakly hierarchical* $(k \geq 2)$ clustering models. The (well known) hierarchical model [17,18,6] is characterized by the absence of overlap, *i.e.*, two clusters are always either disjoint or nested. In the so-called *k*-weakly hierarchical model [2,11,8], the intersection of any $(k + 1)$ clusters should be equal to the intersection of *k* clusters among the $(k + 1)$. When $k = 2$, the 2-weakly hierarchical model, which is simply called weakly hierarchical, has appeared to play a central role in the mathematical theory of clustering: see [1,2,12,3].

Nice properties can be obtained about clusters and clustering models, in the framework of convex structures. Indeed, convexity of clusters has already been considered under different mathematical contexts. One of these contexts underlies scheduling problems and more precisely the task clustering problem (see for example Pecero Sanchez and Trystram [19]). In a close line, several researchers have investigated the so-called pyramidal clustering model, an extension of the hierarchical model, in which all clusters are convex subsets in the sense that they are intervals of some total ranking of the set of objects to

[*] Correspondence to: Université Paris-Dauphine, PSL Research University, Ceremade, 75775 Paris Cedex 16, France.
E-mail addresses: bertrand@ceremade.dauphine.fr (P. Bertrand), jean.diatta@univ-reunion.fr (J. Diatta).

http://dx.doi.org/10.1016/j.dam.2016.12.019
0166-218X/© 2016 Elsevier B.V. All rights reserved.

be clustered (e.g. [13–15]). In what follows, we refer to the notion of convexity in the general abstract sense, as formulated by Van de Vel in [21]: a (abstract) *convexity* on a finite object set S is a collection of subsets of S, called *convex subsets*, containing both the empty set and S, and closed under arbitrary intersections. Thanks to the presence of the ground set S and the closeness under arbitrary intersections, any convexity has a related hull operator that associates each object subset X with the smallest convex subset containing X. The restriction of the hull operator to the object pairs is known as the segment operator. The segment operator, as a symmetric map defined on the set of object pairs, is a so-called interval operator and any interval operator induces a convexity consisting of the collection of all object subsets such that each of these subsets contains the interval between any two of its objects (cf. for example [21]).

The main contributions of the present paper are:

- A necessary and sufficient condition characterizing paired hierarchical convexities in terms of the segment operator;
- Sufficient conditions for an interval convexity to be either hierarchical, paired hierarchical, pyramidal or k-weakly hierarchical;
- An algorithm for computing the interval convexity induced by any given interval operator; this leads to a unified clustering algorithm which encompasses all above mentioned multilevel clustering models, when the input data is an interval operator satisfying the condition specific to the required model.

The paper is organized as follows. Section 2 provides successively a short review of the clustering models considered in this text, and their characterizations in terms of the segment operator. Section 3 is devoted to the study of these clustering models as interval convexities. Section 4 presents an algorithm for computing the interval convexity induced by any interval operator and, finally, a general clustering scheme that unifies the ways to achieve a multilevel clustering with respect to the various clustering models is presented and illustrated in Section 5.

2. Convex hull-based characterizations of multilevel clustering models

In this section, we consider the most known multilevel clustering models dealt with in the literature. We characterize them in terms of their hull operator, after a brief recall of their respective definitions.

Throughout this text, S will denote a finite nonempty object set and $|X|$ the size of any finite set X.

2.1. Multilevel clustering models

The clustering models we consider in the present paper are, from the more specific to the more general, the (strongly) hierarchical, paired hierarchical, pyramidal and k-weakly hierarchical clustering models. In what follows, we review their respective definitions, and provide illustrative examples, constructed, all on the same ground set $\{a, b, c, d, e\}$.

The notion of a *proper intersection* will be used extensively in the following investigation of multilevel clustering models. It is said that two sets X and Y *intersect properly* if each of the sets $(X \setminus Y)$, $(Y \setminus X)$ and $(X \cap Y)$ is nonempty, which is equivalent to assert that $X \cap Y \notin \{X, Y, \emptyset\}$. We also say that X *crosses* Y, or equivalently that Y *crosses* X.

(1) *Hierarchical model.* A collection $\mathscr{C}$ of subsets of S is said to be *(strongly) hierarchical* if no two $\mathscr{C}$-members intersect properly or, in other words, if any two $\mathscr{C}$-members are either disjoint or nested [17,18,6]. Formally, $\mathscr{C}$ is hierarchical if $X \cap Y \in \{X, Y, \emptyset\}$ for all $\mathscr{C}$-members X and Y.

As an illustrative example, the collection $\mathscr{C}_1 = \{\{a, c\}, \{b, e\}, \{b, d, e\}\}$ is clearly hierarchical on the set $\{a, b, c, d, e\}$. Fig. 1(a) depicts the Venn diagram of $\mathscr{C}_1$.

(2) *Paired hierarchical model.* A collection $\mathscr{C}$ of subsets of S is said to be *paired hierarchical* if each $\mathscr{C}$-member intersects properly at most one other [7]. Such a collection $\mathscr{C}$ is named paired hierarchical since there exists a partition $\mathscr{P}$ of $\mathscr{C}$ such that each $\mathscr{P}$-class has at most two members, and any two members in distinct $\mathscr{P}$-classes are either disjoint or nested. One main interest in considering the paired hierarchical clustering model, is that it allows both for a parsimonious collection of clusters, since its maximal size is a linear function of $|S|$, and for overlapping clusters. Clearly, any hierarchical collection is paired hierarchical.

Given the hierarchical collection $\mathscr{C}_1$ defined in paragraph (1), the collection $\mathscr{C}_2 = \mathscr{C}_1 \cup \{\{d, e\}\}$ is an example of paired hierarchical (but not hierarchical) collection on $\{a, b, c, d, e\}$. Fig. 1(b) depicts the Venn diagram of $\mathscr{C}_2$.

(3) *Pyramidal model.* A collection $\mathscr{C}$ of subsets of S is said to be *pyramidal* if there exists a linear order $\preceq$ on S, of which each $\mathscr{C}$-member is an interval [13,14]. By an interval of the order $\preceq$ is meant, as usual, a subset X of S such that for all $x, y \in X$, each element z comprised between x and y in the sense of $\preceq$ (*i.e.* such that $x \preceq z \preceq y$) belongs necessarily to X. It may be noticed that any paired hierarchical collection, hence any hierarchical collection, is pyramidal. The pyramidal clustering model allows for overlapping clusters and for a convenient graphical representation of its clusters by means of its Hasse diagram (i.e. the graph of the covering relation of the inclusion order on $\mathscr{C}$) which is, in this case, a planar graph. However, the maximal size of a pyramidal collection is a quadratic function of $|S|$. Note that in the hierarchical and paired hierarchical cases, the maximal number of clusters is a linear function of $|S|$, and this makes easier the task of interpreting the obtained clusters.

Given the paired hierarchical collection $\mathscr{C}_2$ defined in paragraph (2), $\mathscr{C}_3 = \mathscr{C}_2 \cup \{\{a, c, d\}\}$ is an example of pyramidal (but not paired hierarchical) collection on $\{a, b, c, d, e\}$. Fig. 1(c) provides the Venn diagram of $\mathscr{C}_3$.

Please cite this article in press as: P. Bertrand, J. Diatta, Multilevel clustering models and interval convexities, Discrete Applied Mathematics (2017), http://dx.doi.org/10.1016/j.dam.2016.12.019