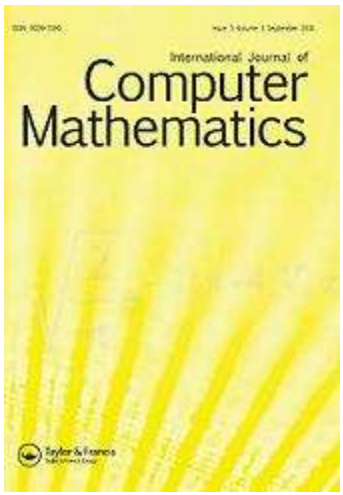
Generalized hard cluster analysis

Wim de Mulder



Generalized hard cluster analysis

Journal:	*International Journal of Computer Mathematics*
Manuscript ID:	GCOM-2010-0812-A
Manuscript Type:	Original Article
Date Submitted by the Author:	18-Oct-2010
Complete List of Authors:	De Mulder, Wim; Ghent University
Keywords:	hard clustering, (m,n)-set, transitivity, matrix representation, equivalent clusterings

International Journal of Computer Mathematics
Vol. 00, No. 00, January 2008, 1–11

RESEARCH ARTICLE

Generalized hard cluster analysis

W. DE MULDER*

Department of Electrical Energy, Systems and Automation, University of Ghent, 9052 Ghent, Belgium

(*Received 00 Month 200x; in final form 00 Month 200x*)

In this paper we generalize the hard clustering paradigm. While in this paradigm a data set is subdivided in disjoint clusters, we allow different clusters to have a nonempty intersection. The concept of hard clustering is then analyzed in this general setting, and we show which specific properties hard clusterings possess in comparison to more general clusterings. We also introduce the concept of equivalent clusterings and show that in case of hard clusterings equivalence and equality coincide. However, if more general clusterings are considered, these two concepts differ and this implies the undesired fact that equivalent clusterings can have different representations in the traditional view on clustering. We show how a matrix representation can solve this representation problem.

Keywords: hard clustering; matrix representation; transitivity; $(m, n)-$set; equivalent clusterings

AMS Subject Classification: CR Category: I.5.3

1. Introduction

Cluster analysis is the partitioning of a data set into subsets (clusters), such that the data elements in each subset are similar to each other and dissimilar to the data elements in other subsets [6, 10]. Similarity and dissimilarity are defined in terms of a distance measure. A clustering algorithm is an algorithm that accomplishes this partitioning [2]. The resulting set of clusters, after applying a clustering algorithm, is called a clustering.

The contribution of this paper is to generalize hard cluster analysis by allowing that a given data element can belong to more than one cluster, while in hard cluster analysis a given data element belongs to one and only one cluster [6]. The most well-known hard clustering algorithm is k-means [12]. An interesting discussion about the evolution of this still widely used algorithm is given in [7].

Furthermore, we extend hard cluster analysis in a different way than fuzzy cluster analysis [1]. Fuzzy cluster analysis also allows that a given data element belongs to more than one cluster, but only to a certain degree between 0 and 1. Moreover, a typical fuzzy clustering algorithm requires that the sum of the degrees equals 1, while in the context of generalized hard cluster analysis we allow that a given data element belongs fully, i.e. to degree 1, to several clusters, although a given data element cannot belong to a cluster to a degree between 0 and 1.

What is the relevance of allowing that a data element can belong to several clusters?

*Email: wim.demulder@ugent.be

ISSN: 0020-7160 print/ISSN 1029-0265 online

DOI: 0020716YYxxxxxxxx
http://www.informaworld.com

We see both theoretical and practical advantages of this generalization. An important theoretical advantage is that a generalized setting allows to derive properties that are specific for hard clusterings, i.e. properties not shared by clusterings for which overlap between clusters is allowed. One practical advantage is that if two different hard clustering algorithms are applied to the same data set, results are mostly different. It is considered to be very hard to find an optimal way to combine these different clusterings [3, 5]. The basic reason for the difficulty of combining different clustering results is the inconsistency between the clusterings, more precisely the fact that while a data element belongs to one cluster according to the first algorithm, it belongs to another cluster according to the second algorithm. This inconsistency may be eliminated by allowing that the considered data element belongs to both clusters in the final clustering. The relaxation of the requirement that each data element belongs to a single cluster thus greatly facilitates the integration of clusterings produced by different hard clustering algorithms, each having its own view on the given data set. The combination of different clusterings is the main topic of the theory of cluster ensembles [4, 9, 11].

To give a specific application of generalized hard cluster analysis, consider the cluster analysis of gene expression data sets (the topic of the cluster analysis of gene expression data sets is discussed in, e.g., [8]). The purpose of gene-based cluster analysis is to group together co-expressed genes (where co-expression is defined in terms of a similarity measure), which indicate co-function and co-regulation. However, it was recently found that the old paradigm that one gene makes one protein does not hold, and that through mechanisms that include alternative splicing, one gene can direct the synthesis of many proteins [13]. This implies that one gene can be involved in several, different processes and thus can have several functions. Generalized hard cluster analysis offers a solution in this case, since it allows that a gene is member of several clusters, and thus this method recognizes that one gene can have several functions. In this application, each cluster could correspond to a certain biological function.
Notice that generalized hard cluster analysis is even more suitable than fuzzy cluster analysis for this application, since a typical fuzzy clustering algorithm assigns each data element to every cluster to a certain degree, but such that the sum of these degrees, for a given element, equals 1. This is not desired in this case, because a gene can fulfill several functions, as discussed above, and it would be inappropriate to say that a gene fulfills a certain function only to a certain degree.

In this work, the focus is on the theoretical side of the mentioned generalization, developing a rigorous framework for generalized hard clustering and deriving specific properties for hard clusterings within this framework. At the same time, some interesting properties that apply to generalized hard clusterings, but not necessarily to hard clusterings, are proven. For example, in section 5 the maximum number of clusters that is needed to arbitrarily subdivide a data set into overlapping clusters is derived.

2. $(m, n) - sets$

In this section we introduce the concept of $(m, n)-$sets that will be used to represent a generalized hard clustering.
Throughout the paper, in particular for the definitions and theorems, it is assumed that a data set is given. The term data set is considered in the most general sense, i.e. a data set is just a finite set $D = \{d_1, \ldots, d_n\}$ containing elements of an

arbitrary nature. Since cluster analysis is about grouping objects, it is supposed that $n \geq 2$.

DEFINITION 2.1 *Given $m, n \in \mathbb{N}$ an (m,n)-set is a set $\Phi = (\Phi_1, \ldots, \Phi_m)$ such that $\Phi_i \in \{(a_1, \ldots, a_n) \,|\, a_j \in \{0,1\},\ \forall 1 \leq j \leq n\}$. Given $\Phi_i \in \Phi$ we denote by Φ_i^j the jth element of Φ_i, i.e. $\Phi_i = (\Phi_i^1, \ldots, \Phi_i^j, \ldots, \Phi_i^n)$. The set of all (m,n)-sets is denoted as $M_{m,n}$.*

Example 1. Given $m = 2, n = 3$, the set $\Phi = \{(1,0,0),(0,0,1)\}$ is an element of $M_{2,3}$.

Our purpose is to use an $(m,n)-$set Φ to represent a generalized hard clustering, where elements of Φ, namely Φ_i, correspond to clusters. Thus in the above example the element $(1,0,0)$ would represent a cluster consisting of d_1, while the element $(0,0,1)$ corresponds to a cluster containing d_3. However, since we require that each element belongs to at least one cluster, the above definition is too general to be used directly as representation for generalized hard clusterings. Consequently, we define the following subset of $M_{m,n}$.

DEFINITION 2.2

$$C_{m,n} = \left\{\Phi = (\Phi_1, \ldots, \Phi_m) \in M_{m,n}\right\} \text{ such that}$$

1. $\sum_{i=1}^{m} \Phi_i^j \geq 1, \ \forall 1 \leq j \leq n$
2. $\Phi_i \neq (0, \ldots, 0), \forall 1 \leq i \leq N$

DEFINITION 2.3 *A clustering is an element of $C_{N,n}$ with $N = n(n-1)/2$ if $n \geq 3$ and $N = 2$ if $n = 2$.*

Remark 1 It is implicitly understood that the term clustering in this work refers to a generalized hard clustering.

Remark 2 As will become clear below, the first requirement ensures that each data element belongs to at least one cluster.

Remark 3 As will become clear below, the second requirement excludes empty clusters. Empty clusters are not interesting to consider, and excluding them avoids the need to give attention to special cases that are related to the presence of empty clusters.

Remark 4 The motivation for the definition and use of N will be given in section 5.

Remark 5 Notice the abstractness of the definition: no reference is made to, e.g., a distance measure.

DEFINITION 2.4

$$H_{m,n} = \left\{\Phi = (\Phi_1, \ldots, \Phi_m) \in C_{m,n}\right\} \text{ such that}$$

1. $\sum_{i=1}^{m} \Phi_i^j = 1, \ \forall 1 \leq j \leq n$
2. $\Phi_i \neq (0, \ldots, 0), \forall 1 \leq i \leq N$

Remark 6 The set $H_{m,n}$ thus contains all (m,n)-sets for which for a given j there

exists one and only one i such that $\Phi_i^j = 1$, and such that no Φ_i equals $(0, \ldots, 0)$.

DEFINITION 2.5 *A* hard *clustering is an element of* $H_{N,n}$.

If no confusion can arise, the shorter notations C and H are used instead of $C_{N,n}$ resp. $H_{N,n}$.

DEFINITION 2.6 *Given* $\Phi = (\Phi_1, \ldots, \Phi_N) \in C$ *we define*

$$\mathcal{F}(\Phi_i) = \{d_j \in D \,|\, \Phi_i^j = 1\}$$

for $i = 1, \ldots, N$. *The sets* $\mathcal{F}(\Phi_i)$ *are called the clusters of* Φ.

Remark 7 The function $\mathcal{F}$ thus links an element of a clustering $\Phi \in C$ to a subset of the data set.

Example 2. Given $D = \{d_1, d_2, d_3\}$ the following is a clustering: $\Phi = ((1,1,0),(0,1,1),(0,0,1))$ and $\{\{d_1, d_2\}, \{d_2, d_3\}, \{d_3\}\}$ is the corresponding set of clusters. Notice that $\Phi \in C_{N,n}$ with $n = 3$, but $\Phi \notin H_{N,n}$, since $\sum_{i=1}^{3} \Phi_i^2 = 2 \neq 1$ which contradicts definition 2.4.
From definition 2.6 and example 2 it is seen that each element $\Phi_i \in \Phi$ corresponds to a cluster. This means that the representation of clusterings by (m, n)-sets is equivalent to the classical representation of a clustering where a clustering consists of groups of data elements. Below it will be shown that this representation is suitable for hard clusterings, but inappropriate to represent generalized hard clusterings. In section 5 an alternative representation will be presented that is more convenient for generalized hard clusterings.

DEFINITION 2.7 *Given* $\Lambda \subseteq \{1, \ldots, n\}$ *and* $\Phi \in C_{N,n}$ *we define* $\Phi_i^\Lambda = 1$ *if and only if* $\Phi_i^j = 1,\ \forall j \in \Lambda$ *and* $\Phi_i^\Lambda = 0$ *if and only if* $\Phi_i^j = 0,\ \forall j \in \Lambda$.

DEFINITION 2.8 *Given* $\Lambda \subseteq \{1, \ldots, n\}$ *and* $\Phi \in C_{N,n}$ *we define* $\Phi^\Lambda = \{\Phi_i \in \Phi \,|\, \Phi_i^\Lambda = 1\}$.

Remark 8 From the above definition it follows that if $\Phi_i \in \Phi^\Lambda$, then $\{d_j \in D \,|\, j \in \Lambda\} \subseteq \mathcal{F}(\Phi_i)$. Thus $\Phi_i \in \Phi^\Lambda$ corresponds to a cluster that contains, at least, all data elements d_j with $j \in \Lambda$.

Example 3. Given the results of example 2 we have, for example, $\Phi^{\{2,3\}} = \{(0,1,1)\}$. These are all the clusters, in this case only one, such that *both* d_2 and d_3 belong to them. Another example is $\Phi^{\{2\}} = \{(1,1,0),(0,1,1)\}$.

THEOREM 2.9 $C_{N,n} = \{\Phi \in M_{N,n} \,|\, \Phi^{\{j\}} \neq \emptyset,\ \forall 1 \leq j \leq n, \Phi_i \neq (0, \ldots, 0), \forall 1 \leq i \leq N\}$.

Proof Let $\Phi \in C_{N,n}$. From definition 2.1 and 2.2 it follows that $\Phi_i^j \in \{0, 1\}$ and $\sum_{i=1}^{N} \Phi_i^j \geq 1,\ \forall 1 \leq j \leq n$. This is equivalent to saying that given $1 \leq j \leq n$ there exists a $1 \leq i \leq N$ for which $\Phi_i^j = 1$. By definition 2.8 this is equivalent to $\Phi^{\{j\}} \neq \emptyset, \forall 1 \leq j \leq n$.
Conversely, if $\Phi \in M_{N,n}$ and $\Phi^{\{j\}} \neq \emptyset$ then, by definition 2.8, there exists a $1 \leq i \leq N$ such that $\Phi_i^j = 1$ and thus $\sum_{i=1}^{N} \Phi_i^j \geq 1$. Since this holds for all $1 \leq j \leq n$ it follows from definition 2.2 that $\Phi \in C_{N,n}$. ■

Discussion. The above theorem states that, loosely speaking, a clustering is a subdivision of a data set into clusters such that each data element belongs to at least one cluster. The theorem could also be interpreted as stating that $(m, n)-$sets are the right concept for representing generalized hard clusterings, where each data element belongs to at least one cluster. However, in the next section we develop

the concept of equivalent clusterings, and in light of this concept we will conclude that $(m,n)-$sets fall short in providing an adequate representation of generalized hard clusterings.

3. Equivalent generalized hard clusterings

Since clustering is about grouping similar objects, we should describe two clusterings in which the same objects are considered as similar, as being equal. However, it is possible that two such clusterings have a different representation in terms of (m,n)-sets, as will be shown in a moment. Thus the fact that two clusterings represent the same grouping of objects cannot be expressed by stating that their representation by (m,n)-sets are equal. To resolve this, we introduce the concept of equivalent clusterings.

DEFINITION 3.1 $\Phi \equiv \Psi$ *if the following holds:* $\Phi^{\{i,j\}} = \emptyset \Leftrightarrow \Psi^{\{i,j\}} = \emptyset$, $\forall i,j \in \{1,\ldots,n\}$.

Remark 1 The interpretation of the definition is that two equivalent clusterings represent the same grouping of objects. For example, let d_1, d_2 and d_3 be elements of a cluster in $\Phi \equiv \Psi$, i.e. $\exists\, \Phi_i$ such that $\{d_1,d_2,d_3\} \subseteq \mathcal{F}(\Phi_i)$. By definitions 2.6 and 2.7 this implies that $\Phi_i^{\{1,2,3\}} = 1$. In particular, $\Phi_i^{\{1,2\}} = 1$ and thus, by definition 2.8, $\Phi^{\{1,2\}} \neq \emptyset$. This implies, by the above definition, that $\Psi^{\{1,2\}} \neq \emptyset$. In the same way we find that $\Psi^{\{1,3\}} \neq \emptyset$ and $\Psi^{\{2,3\}} \neq \emptyset$. Thus in the clustering Ψ it holds that d_1 and d_2 are grouped together, as well as d_1 and d_3, and d_2 and d_3, although it is not required that there is one cluster in Ψ such that all these three elements belong to it. It is however clear that adding a cluster to Ψ consisting of d_1, d_2 and d_3 should not alter the information contained in Ψ, and this precisely means that equivalent clusterings should be considered as *equivalent*, in the sense of representing the same subdivision of the given data set.

Example 4. Consider $\Phi = ((1,1,0),(1,0,1),(0,1,1))$ and $\Psi = ((1,1,1),(0,0,1),(1,0,0))$. Then it is easily verified that $\Phi \equiv \Psi$. It is noticed however that $\Phi \neq \Psi$.

DEFINITION 3.2 $R_\Phi = \{\Psi \in C \,|\, \Psi \equiv \Phi\}$.

Given a finite set A the notation $|A|$ is used to denote its number of elements.

THEOREM 3.3 $\Phi \in H \Rightarrow R_\Phi \cap H = \{\Phi\}$.

Proof Given is $\Phi = (\Phi_1,\ldots,\Phi_N) \in H$. Since definition 3.1 assures that $\Phi \equiv \Phi$ it follows that $\Phi \in R_\Phi \cap H$. Now suppose that there exists a $\Psi = (\Psi_1,\ldots,\Psi_N) \in H \cap R_\Phi$ with $\Phi \neq \Psi$. Thus there exists a i such that either $\Psi_i \in \Psi \setminus \Phi$ or $\Phi_i \in \Phi \setminus \Psi$. Since both cases are entirely similar we consider only the case where $\Psi_i \in \Psi \setminus \Phi$. Let $A_1 = \{j \in \{1,\ldots,n\} \,|\, \Psi_i^j = 1\}$ and $A_0 = \{1,\ldots,n\} \setminus A_1$. Since $\Psi_i \notin \Phi$ we either have that $\Phi_i^{A_1} \neq 1, \forall\, 1 \leq i \leq N$ or $\Phi_i^{A_0} \neq 0, \forall\, 1 \leq i \leq N$.

Case 1: $\Phi_i^{A_1} \neq 1, \forall\, 1 \leq i \leq N$.
This implies that $|A_1| \geq 1$, since otherwise we would have that $(0,\ldots,0) \in \Psi$. This is contradictory to $\Psi \in H$, by definition 2.4.
Thus $|A_1| \geq 1$, let us say $A_1 = \{k_1,\ldots,k_s\}$. Since $\Phi \in H$ it follows from definition 2.4 that $\sum_{m=1}^N \Phi_m^{k_1} = 1$ and thus there exists exactly one j for which $\Phi_j^{k_1} = 1$. Now since $\Phi \equiv \Psi$ and $\Psi_i^{A_1} = 1$ it follows that $\Phi^{\{k_1,k_2\}} \neq \emptyset$, thus there exists a j_2 such that $\Phi_{j_2}^{\{k_1,k_2\}} = 1$. Again applying the condition $\sum_{m=1}^N \Phi_m^{k_1} = 1$ results in the

fact that $j = j_2$, thus $\Phi_j^{\{k_1,k_2\}} = 1$. This can now be repeated for $k_3, \ldots, k_s$ from which it is found that $\Phi_j^{A_1} = 1$, which is contradictory to $\Phi_i^{A_1} \neq 1, \forall 1 \leq i \leq N$.

Case 2: the case $\Phi_i^{A_0} \neq 0, \forall 1 \leq i \leq N$ is handled entirely analogous as above. ■

Discussion. The above theorem says that when attention is restricted to hard clustering the concepts of equivalence and equality are the same. Thus in case of hard clustering an (m, n)-set provides a *unique* representation for a given set of equivalent clusterings. Since equivalent clusterings represent the same subdivision of a given data set, the representation by (m, n)-sets is thus well-suited for the special case of hard clustering.
However, extending hard clusterings to generalized hard clusterings it is no longer true that equivalence and equality in terms of (m, n)-sets are the same concepts, as example 4 illustrates. In the section 5 we show how generalized hard clusterings can be represented in matrix form and that in terms of this matrix representation equivalence and equality are the same concept.
In the next section we define the concept reflective transitivity and show that this property is characteristic for hard clusterings.

4. Transitivity

DEFINITION 4.1 $\Phi \in C$ *is transitive if the following holds:* $\{d_{i_1}, d_{i_2}\} \subseteq \mathcal{F}(\Phi_i), \{d_{i_2}, d_{i_3}\} \subseteq \mathcal{F}(\Phi_j) \Rightarrow \exists\, \Psi \in R_\Phi, \exists k \in \{1, \ldots, N\} : \{d_{i_1}, d_{i_2}, d_{i_3}\} \subseteq \mathcal{F}(\Psi_k)$.

Remark 1 The above definition states that, given any three data elements $d_{i_1}, d_{i_2}, d_{i_3}$, a clustering is transitive if it holds that d_{i_1} and d_{i_2} belong to the same cluster and the same is true for d_{i_2} and d_{i_3}, then all three elements belong to the same cluster in a equivalent clustering. Thus the definition does not require that the cluster, or clusters, to which these three elements belong, is an element of the given clustering, but only that it is an element of an equivalent clustering.

The next theorem states the above definition in another form, without making reference to $\mathcal{F}$.

THEOREM 4.2 $\Phi \in C$ *is transitive if and only if the following holds:* $\Phi^{\{i_1,i_2\}} \neq \emptyset, \Phi^{\{i_2,i_3\}} \neq \emptyset \Rightarrow \exists\, \Psi \in R_\Phi : \Psi^{\{i_1,i_2,i_3\}} \neq \emptyset$.

Proof Suppose that $\Phi \in C$ is transitive and $\Phi^{\{i_1,i_2\}} \neq \emptyset, \Phi^{\{i_2,i_3\}} \neq \emptyset$. Then by definition 2.8 there exist a i and j such that $\Phi_i^{\{i_1,i_2\}} = 1, \Phi_j^{\{i_2,i_3\}} = 1$ and by definition 2.6 we can write that $\{d_{i_1}, d_{i_2}\} \subseteq \mathcal{F}(\Phi_i), \{d_{i_2}, d_{i_3}\} \subseteq \mathcal{F}(\Phi_j)$. By definition 4.1 there exists a $\Psi \in R_\Phi, \exists k \in \{1, \ldots, N\}$ such that $\{d_{i_1}, d_{i_2}, d_{i_3}\} \subseteq \mathcal{F}(\Psi_k)$. Applying definition 2.6 again this means that $\Psi_k^{i_1} = \Psi_k^{i_2} = \Psi_k^{i_3} = 1$ and thus $\Psi^{\{i_1,i_2,i_3\}} \neq \emptyset$.
Conversely, suppose that $\Phi^{\{i_1,i_2\}} \neq \emptyset, \Phi^{\{i_2,i_3\}} \neq \emptyset \Rightarrow \exists\, \Psi \in R_\Phi : \Psi^{\{i_1,i_2,i_3\}} \neq \emptyset$. Suppose that $\{d_{i_1}, d_{i_2}\} \subseteq \mathcal{F}(\Phi_i), \{d_{i_2}, d_{i_3}\} \subseteq \mathcal{F}(\Phi_j)$. By definitions 2.6 and 2.8, this is equivalent to saying that $\Phi^{\{i_1,i_2\}} \neq \emptyset, \Phi^{\{i_2,i_3\}} \neq \emptyset$. By assumption it then follows that there exists a $\Psi \in R_\Phi$ for which $\Psi^{\{i_1,i_2,i_3\}} \neq \emptyset$ and definition 2.8 then guarantees that there exists a $k \in \{1, \ldots, N\}$ such that $\Psi_k^{\{i_1,i_2,i_3\}} = 1$ or in other words: $\{d_{i_1}, d_{i_2}, d_{i_3}\} \subseteq \mathcal{F}(\Psi_k)$ from which we conclude that Φ is transitive. ■

DEFINITION 4.3 *A clustering* $\Phi \in C$ *is reflective transitive if the following holds:* $\{d_{i_1}, d_{i_2}\} \subseteq \mathcal{F}(\Phi_i), \{d_{i_2}, d_{i_3}\} \subseteq \mathcal{F}(\Phi_j) \Rightarrow \exists k \in \{1, \ldots, N\} : \{d_{i_1}, d_{i_2}, d_{i_3}\} \subseteq \mathcal{F}(\Phi_k)$.

Remark 2 Notice that unlike definition 4.1, the above definition requires that the

elements d_{i_1}, d_{i_2} and d_{i_3} belong to a cluster of the given clustering, and not just to a cluster of an equivalent clustering.

THEOREM 4.4 *$\Phi \in C$ is reflective transitive if and only if the following holds: $\Phi^{\{i_1,i_2\}} \neq \emptyset, \Phi^{\{i_2,i_3\}} \neq \emptyset \Rightarrow \Phi^{\{i_1,i_2,i_3\}} \neq \emptyset$.*

Proof The proof is entirely analogous to the proof of theorem 4.2. ■

THEOREM 4.5 *If $\Phi \in H$ then Φ is reflective transitive.*

Proof Suppose that $\Phi^{\{i_1,i_2\}} \neq \emptyset, \Phi^{\{i_2,i_3\}} \neq \emptyset$. By definition 2.8 there exists a i, j such that $\Phi_i^{\{i_1,i_2\}} = \Phi_j^{\{i_2,i_3\}} = 1$. Since $\Phi \in H$ the condition $\sum_{m=1}^{N} \Phi_m^{i_2} = 1$ holds, which is only possible if $i = j$. Thus $\Phi_i^{\{i_1,i_2,i_3\}} = 1$ which proves the proposition. ■

Remark 3 The above theorem and the corresponding proof ensure that, in case of hard clustering, if d_1 and d_2 belong to the same cluster and the same is true for d_2 and d_3, then d_1, d_2 and d_3 belong to the same cluster. It is interesting to notice that this transitive property does not necessarily hold for a generalized hard clustering. This is seen in example 2 where $\Phi^{\{1,2\}} \neq \emptyset, \Phi^{\{2,3\}} \neq \emptyset$, but $\Phi^{\{1,2,3\}} = \emptyset$.

DEFINITION 4.6 *Given $\Phi \in C$ we define: $\mathcal{F}(\Phi) = \{\mathcal{F}(\Phi_1), \ldots, \mathcal{F}(\Phi_N)\}$.*

DEFINITION 4.7 $\mathcal{F}_{\mathbf{C}} = \{\mathcal{F}(\Phi) \,|\, \Phi \in C\}$.

Remark 4 It is easily seen that $\mathcal{F} : C \Rightarrow \mathcal{F}_{\mathbf{C}}$ is invertible.

DEFINITION 4.8 *Given $A \subseteq D$ we define its index set, denoted I_A, as $I_A = \{1 \leq i \leq n \,|\, d_i \in A\}$.*

In fact, definition 4.7 defines the set of all possible subdivisions of a given data set. This is now proved.

THEOREM 4.9 $\mathcal{F}_{\mathbf{C}} = \left\{\{F_1, \ldots, F_N\} \,|\, F_i \subseteq D,\ \cup_{i=1}^{N} I_{F_i} = \{1, \ldots, n\}\right\}$.

Proof Consider any $\Phi = (\Phi_1, \ldots, \Phi_N) \in C$ and $1 \leq j \leq n$. From definition 2.6 we can already conclude that $\mathcal{F}(\Phi_i) \subseteq D$. Since $\Phi \in C$ we know that $\sum_{i=1}^{N} \Phi_i^j \geq 1$. Thus there exists at least one i such that $\Phi_i^j = 1$ and thus $d_j \in \mathcal{F}(\Phi_i)$ by definition 2.6, which implies that $j \in I_{\mathcal{F}(\Phi_i)}$. This allows to write that $\{1, \ldots, n\} \subseteq \cup_{i=1}^{N} I_{\mathcal{F}(\Phi_i)}$. Furthermore, since $\mathcal{F}(\Phi_i) \subseteq D$ it follows that $I_{\mathcal{F}(\Phi_i)} \subseteq \{1, \ldots, n\}$ and thus $\cup_{i=1}^{N} I_{\mathcal{F}(\Phi_i)} \subseteq \{1, \ldots, n\}$. It then that $\cup_{i=1}^{N} I_{\mathcal{F}(\Phi_i)} = \{1, \ldots, n\}$ and thus $\mathcal{F}(\Phi) \in \mathcal{F}_{\mathbf{C}}$.
Conversely, given $\{F_1, \ldots, F_N\}$ for which $F_i \subseteq D$ and $\cup_{i=1}^{N} I_{F_i} = \{1, \ldots, n\}$ define $\Phi = (\Phi_1, \ldots, \Phi_N)$ with $\Phi_i^j = 1$ if $d_j \in F_i$ and $\Phi_i^j = 0$ otherwise, for $j = 1, \ldots, n$. It is then easily seen that $\mathcal{F}(\Phi) = \{F_1, \ldots, F_N\}$. Furthermore given any $m \in \{1, \ldots, n\} = \cup_{i=1}^{N} I_{F_i}$ there exists a i such that $m \in I_{F_i}$ or equivalently $d_m \in F_i$, and by construction of Φ we thus have that $\Phi_i^m = 1$. From this it follows that $\sum_{i=1}^{N} \Phi_i^m \geq 1$ and consequently $\Phi \in C$. Thus $\mathcal{F}(\Phi) = \{F_1, \ldots, F_N\} \in \mathcal{F}_{\mathbf{C}}$ by definition 4.7. ■

DEFINITION 4.10 $\mathcal{F}_{\mathbf{H}} = \{\mathcal{F}(\Phi) \,|\, \Phi \in H\}$.

The above definition represents all possible subdivisions of a given data set in *disjoint* classes (clusters), as ensured by the following theorem.

THEOREM 4.11 $\mathcal{F}_{\mathbf{H}} = \left\{\{F_1, \ldots, F_N\} \,|\, F_i \subseteq D,\ \cup_{i=1}^{N} I_{F_i} = \{1, \ldots, n\},\ F_i \cap F_j = \emptyset \text{ if } i \neq j\right\}$.

Proof Consider any $\Phi = (\Phi_1, \ldots, \Phi_N) \in H$. By the previous theorem it remains only to proof that $\mathcal{F}(\Phi_i) \cap \mathcal{F}(\Phi_j) = \emptyset$ if $i \neq j$. Suppose that $d_k \in \mathcal{F}(\Phi_i) \cap \mathcal{F}(\Phi_j), i \neq j$. Then $\Phi_i^k = \Phi_j^k = 1$ by definition 2.6. This is only compatible with the properties that $\sum_{p=1}^{N} \Phi_p^k = 1$ and $\Phi_p^k = 0$ or 1, if $i = j$ which is a contradiction. Thus $\mathcal{F}(\Phi_i) \cap \mathcal{F}(\Phi_j) = \emptyset$ if $i \neq j$. Conversely, given $\{F_1, \ldots, F_N\}$ for which $F_i \subseteq D, \cup_{i=1}^{N} I_{F_i} = \{1, \ldots, n\}$ and $F_i \cap F_j = \emptyset$ if $i \neq j$ define: $\Phi = (\Phi_1, \ldots, \Phi_N)$ with $\Phi_i^j = 1$ if $d_j \in F_i$ and $\Phi_i^j = 0$ otherwise, for $j = 1, \ldots, n$. It is easy to check that $\mathcal{F}(\Phi) = \{F_1, \ldots, F_N\}$.
It remains to check that $\Phi \in H$. Consider any $m \in \{1, \ldots, n\} = \cup_{i=1}^{N} I_{F_i}$. Then there exists a i such that $m \in I_{F_i}$ which implies that $\Phi_i^m = 1$. Suppose that there is a i_2 for which $\Phi_{i_2}^m = 1$. This implies that $d_m \in F_i \cap F_{i_2}$ which is only possible if $i = i_2$. Thus $\sum_{i=1}^{N} \Phi_i^m = 1$ implying that $\Phi \in H$. ■

Discussion. We conclude from theorems 4.9 and 4.11 that a generalized hard clustering can represent any subdivision of a given data set, allowing overlap, while a hard clustering can only subdivide a given data set in disjoint clusters. Thus our abstract definition 2.4 of a hard clustering conforms the common notion that hard clusters do not overlap.

5. Matrix representation of a clustering

Given $A \in \mathbb{R}^{n \times n}$ and $\emptyset \subset \Lambda, \Gamma \subseteq \{1, \ldots, n\}$ we use the notation $A(\Lambda, \Gamma)$ to denote the submatrix of A containing the elements $A(i, j)$ for which $i \in \Lambda, j \in \Gamma$. The notation $A(i, j)$ will be used as shorthand for $A(\{i\}, \{j\})$.

DEFINITION 5.1 *Given $\Phi \in C$ we define its matrix representation, denoted as $M(\Phi)$, as*

$$\begin{aligned} M(\Phi)(i,j) = 1 &\Leftrightarrow \Phi^{\{i,j\}} \neq \emptyset \\ = 0 &\quad \textit{otherwise} \end{aligned}$$

Remark 1 Thus if for a given clustering Φ we have that $M(\Phi)(i, j) = 1$, then it holds that there exists a cluster in this clustering to which both d_i and d_j belong.

THEOREM 5.2 $\Phi \equiv \Psi \Rightarrow M(\Phi) = M(\Psi)$.

Proof By definition 4.1: $\Phi \equiv \Psi$ if $\Phi^{\{i,j\}} = \emptyset \Leftrightarrow \Psi^{\{i,j\}} = \emptyset$ for all $i, j \in \{1, \ldots, n\}$. From this it is easily seen that $M(\Phi) = M(\Psi)$. ■

THEOREM 5.3 $\Phi \not\equiv \Psi \Rightarrow M(\Phi) \neq M(\Psi)$.

Proof Given is that $\Phi \not\equiv \Psi$. Thus there exists a i and j such that $\Phi^{\{i,j\}} = \emptyset$ and $\Psi^{\{i,j\}} \neq \emptyset$ or $\Phi^{\{i,j\}} \neq \emptyset$ and $\Psi^{\{i,j\}} = \emptyset$. Without loss of generality we suppose that $\Phi^{\{i,j\}} = \emptyset$ and $\Psi^{\{i,j\}} \neq \emptyset$. Then $M(\Phi)(i, j) = 0$ and $M(\Psi)(i, j) = 1$, and thus $M(\Phi) \neq M(\Psi)$. ■

Theorems 5.2 and 5.3 taken together give: $M(\Phi) = M(\Psi) \Leftrightarrow \Phi \equiv \Psi$. Thus while an (m, n)-set gives a unique representation for a set of *equal* clusterings, a matrix representation gives a unique representation for a set of *equivalent* clusterings. Since equivalent clusterings represent the same subdivision of a data set, the matrix representation should be preferred over the representation by (m, n)-sets when dealing with generalized hard clusterings.

DEFINITION 5.4 $M_C^{n \times n} = \{A \in \mathbb{R}^{n \times n} \mid A(i, j) \in \{0, 1\}, A(i, j) = A(j, i), A(i, i) = 1, 1 \leq i, j \leq n\}$

Notice that from definition 5.1 it easily follows that $M(C) \subseteq M_C^{n\times n}$ (we remind the reader that C is used as shorthand for $C_{N,n}$). More is true.

THEOREM 5.5 *The mapping $M : C \to M_C^{n\times n}$ is surjective.*

Proof Given $A \in M_C^{n\times n}$ we have to proof that there exists a $\Phi \in C$ such that $M(\Phi) = A$. We prove this by induction on n.
First suppose that $n = 2$. Then either $A(1,2) = A(2,1) = 1$ or $A(1,2) = A(2,1) = 0$. In the first case define $\Phi = \{(1,1),(1,0)\}$ and in the second case define $\Phi = \{(1,0),(0,1)\}$. Then $\Phi \in C_{2,2}$ and $M(\Phi) = A$.
Suppose that the theorem holds for $n = 2,\ldots,k-1$. Consider now $n = k$ and an arbitrary $A \in M_C^{k\times k}$. We are in search for a $\Phi = (\Phi_1,\ldots,\Phi_K) \in C_{K,k}$ such that $M(\Phi) = A$, where $K = k(k-1)/2$. By induction there exists a $\hat{\Phi} = (\hat{\Phi}_1,\ldots,\hat{\Phi}_{K'}) \in C_{K',k-1}$ with $K' = (k-1)(k-2)/2$ for which $M(\hat{\Phi}) = A(\{1,\ldots,k-1\},\{1,\ldots,k-1\})$. Denote $\hat{\Phi}_i = (\hat{\Phi}_i^1,\ldots,\hat{\Phi}_i^{k-1}), i = 1,\ldots,K'$.

Case 1: $A(i,k) = A(k,i) = 0\ \forall\, 1 \le i \le k-1$. Define $\Phi_i = (\hat{\Phi}_i^1,\ldots,\hat{\Phi}_i^{k-1},0)$ for $1 \le i \le K', \Phi_{K'+1} = (0,\ldots,0,1)$ and $\Phi_i = (1,0,\ldots,0)$ if $K'+1 < i \le K$. Then it is easily checked that $\Phi \in C_{K,k}$ and $M(\Phi) = A$.

Case 2: $\exists\, 1 \le i \le k-1$ for which $A(i,k) = A(k,i) = 1$. Let $I = \{i \in \{1,\ldots,k-1\} \mid A(k,i) = 1\}$. Then we define $\Phi_i = (\hat{\Phi}_i^1,\ldots,\hat{\Phi}_i^{k-1},0)$ if $1 \le i \le K'$ and $\Phi_{K'+i} = 0$ for $i \in \{1,\ldots,k-1\} \setminus I$; finally we define for $i \in I$: $\Phi^i_{K'+i} = \Phi^k_{K'+i} = 1$ and $\Phi^j_{K'+i} = 0$ if $j \notin \{i,k\}$. Notice that $K'+i \le K'+(k-1) = (k-1)(k-2)/2+k-1 = k(k-1)/2 = K$, which is a first requirement to have that $\Phi \in C_{K,k}$. The other requirements are easily checked and we conclude that $\Phi \in C_{K,k}$. We now prove that $M(\Phi) = A$ which amounts to show that $A(i,j) = 1 \Leftrightarrow \Phi^{\{i,j\}} \neq \emptyset$ and $A(i,j) = 0$ otherwise. If $\{i,j\} \subseteq \{1,\ldots,k-1\}$ this is already true by induction and by the construction of Φ from $\hat{\Phi}$. So suppose that $i = k$ and that $\Phi^{\{k,j\}} \neq \emptyset$ which implies the existence of a $1 \le m \le K$ such that $\Phi^j_m = \Phi^k_m = 1$. By construction of Φ it has to be that Φ^j_m is of the form $\Phi^j_{K'+j}$ with $j \in I$. From the definition of I it then follows that $A(k,j) = 1$. If $j = k$ then by definition 5.4 we have that $A(k,k) = 1$. Finally, if $\Phi^{\{k,j\}} = \emptyset$ it follows from the construction of Φ that $j \notin I$ which implies that $A(k,j) = 0$. ■

Remark 2 The proof of the above theorem shows why we have chosen $N = n(n-1)/2$ if $n \ge 3$: this ensures that the mapping $M : C \to M_C^{n\times n}$ is surjective. The case where $n = 2$ is rather a basic case and the proof also shows why in this case $N = 2$ is chosen.
In other words: since clusterings from C represent any subdivision of a given data set (see theorem 4.9) and since the above mapping $M : C_{N,n} \to M_C^{n\times n}$ is surjective if $N = n(n-1)/2$ for $n \ge 3$ and $N = 2$ i $n = 2$, this ensures that one needs never more than N clusters to arbitrary subdivide a given data set.

DEFINITION 5.6 $M_H^{n\times n} = \{A \in M_C^{n\times n} \mid A(i,j) = 1, A(j,k) = 1 \Rightarrow A(i,k) = 1\}$

THEOREM 5.7 *The mapping $M : H \to M_H^{n\times n}$ is injective.*

Proof Given $\Phi,\Psi \in H$ with $\Phi \neq \Psi$ we have to proof that $M(\Phi) \neq M(\Psi)$. Since $\Phi,\Psi \in H$ and $\Phi \neq \Psi$ it follows from theorem 3.3 that $\Psi \notin R_\Phi$, thus $\Phi \not\equiv \Psi$. Thus there exist i,j with $i \neq j$ such that $\Phi^{\{i,j\}} \neq \emptyset, \Psi^{\{i,j\}} = \emptyset$ (or vice versa). By definition of the mapping M it then follows that $M(\Phi)(i,j) = 1, M(\Psi)(i,j) = 0$ and thus in particular: $M(\Phi) \neq M(\Psi)$. ■

THEOREM 5.8 *The mapping $M : H \to M_H^{n\times n}$ is surjective.*

Proof We prove this again by induction on n. The case $n = 2$ is the same as

in theorem 5.5, so suppose the theorem holds up to $k-1, k \geq 2$. Consider now $n = k$ and an arbitrary $A \in M_H^{k\times k}$. It is asked to give a $\Phi = (\Phi_1, \ldots, \Phi_K) \in H_{K,k}$ such that $M(\Phi) = A$. Define $K' = (k-1)(k-2)/2$. By induction there exists a $\hat{\Phi} = (\hat{\Phi}_1, \ldots, \hat{\Phi}_{K'}) \in H_{K',k-1}$ for which $M(\hat{\Phi}) = A(\{1, \ldots, k-1\}, \{1, \ldots, k-1\})$.

Case 1: $A(i,k) = A(k,i) = 0,\ \forall 1 \leq i \leq k-1$. This case can be handled in the same way as in theorem 5.5, giving $\Phi_i = (\hat{\Phi}_i^1, \ldots, \hat{\Phi}_i^{k-1}, 0)$ for $i = 1, \ldots, K', \Phi_{K'+1} = (0, \ldots, 0, 1)$ and $\Phi_i = (1, 0, \ldots, 0)$ for $K'+1 < i \leq K$. It is seen that $\sum_{i=1}^K \Phi_i^k = 1$. By induction and by the construction of Φ from $\hat{\Phi}$: $\sum_{i=1}^{K'} \Phi_i^j = 1,\ \forall 1 \leq j \leq k-1$, and since $\Phi_p^j = 0$ for $p = K'+1, \ldots, K$ this is equivalent to $\sum_{i=1}^K \Phi_i^j = 1,\ \forall 1 \leq j \leq k-1$ implying that $\Phi \in H_{K,k}$.

Case 2: $\exists\, 1 \leq i \leq k-1$ for which $A(i,k) = A(k,i) = 1$. Let $I = \{i \in \{1, \ldots, k-1\} \mid A(k,i) = 1\}$. Consider any $\alpha \in I$. Since by induction $\sum_{j=1}^{K'} \hat{\Phi}_j^\alpha = 1$ there exists exactly one $1 \leq m \leq K'$ such that $\hat{\Phi}_m^\alpha = 1$. Define $\Phi_m = (\hat{\Phi}_m^1, \ldots, \hat{\Phi}_m^{k-1}, 1)$. Furthermore define $\Phi_j = (\hat{\Phi}_j^1, \ldots, \hat{\Phi}_j^{k-1}, 0)$ for $1 \leq j \neq m \leq K'$ and $\Phi_j = 0$ if $K' < j \leq K$. We now prove that $\Phi \in H_{K,k}$. By induction and by the construction of Φ from $\hat{\Phi}$ it follows, in the same way as in case 1, that $\sum_{j=1}^K \Phi_j^i = 1$ if $1 \leq i \leq k-1$. The fact that $\Phi_j^k = 1$ if and only if $j = m$ implies that $\sum_{j=1}^K \Phi_j^k = 1$.
Finally we prove that $M(\Phi) = A$. Thus consider i, j for which $\Phi^{\{i,j\}} \neq \emptyset$. If $i, j \in \{1, \ldots, k-1\}$ induction ensures that $M(\Phi)(i,j) = 1$. Suppose that $i = k$, i.e. there exists a t such that $\Phi_t^k = \Phi_t^j = 1$. If $j = k$ we have that $A(k,k) = 1$ by definition of M_H, so suppose that $j \neq k$. By construction we know that $\Phi_m^\alpha = \Phi_m^k = 1$ and thus that $\Phi^{\{\alpha,k\}} \neq \emptyset$. Since $\Phi^{\{\alpha,k\}} \neq \emptyset$ and $\Phi^{\{j,k\}} \neq \emptyset$ theorem 4.5 implies that $\Phi^{\{\alpha,j,k\}} \neq \emptyset$. Thus $\Phi^{\{j,\alpha\}} \neq \emptyset$ and since $1 \leq j, \alpha \leq k-1$ it follows from induction that $A(j,\alpha) = 1$. Together with $A(k,\alpha) = 1$ and the transitive property of M_H it follows that $A(k,j) = 1$. Finally, if $\Phi^{\{k,j\}} = \emptyset$ it follows from the construction of Φ that $j \notin I$ which implies that $A(k,j) = 0$. ■

Discussion. Theorem 5.7 and 5.8, together with definition 5.6, indicate that transitivity is the distinctive property of hard clusterings compared to general clusterings. Theorem 4.11 showed that the distinctive property of hard clusterings Φ is given by $\mathcal{F}(\Phi_i) \cap \mathcal{F}(\Phi_j) = \emptyset$ if $i \neq j$; informally this can also be stated as the absence of overlap. Definition 2.4 indicates that the fact that each data element belongs to exactly one cluster is also the distinctive property of hard clusterings. From these three considerations it can be concluded that absence of overlap, transitivity and the property that each data element belongs to exactly one cluster, are equivalent properties.

6. Conclusion and future work

We generalized the hard clustering paradigm by allowing that a data element can belong to more than one cluster. It is shown that generalized hard clusterings can be represented by (m,n)-sets, a new concept that we discussed in this paper, and by matrices. The concept of equivalent generalized hard clusterings was defined and it was shown that in light of this concept the representation by matrices should be preferred.
Another new concept we introduced was that of transitive clusterings. Using the paradigm of generalized hard clusterings, it was demonstrated that transitivity, absence of overlap between clusters and the property that each data element belongs to exactly one cluster are synonymous characteristics of hard clusterings.
An interesting and important open question is how to deduce the minimal num-

ber of clusters, given a generalized hard clustering. This is a relevant question, since it is possible that redundant or useless information is present in a generalized hard clustering. For example, consider the generalized hard clustering $\Phi = ((1,1,0),(0,1,0),(0,0,1))$. This corresponds to three clusters where the first cluster represents the information that d_1 and d_2 are similar to each other, the second cluster states that d_2 is similar to itself and the third cluster states that d_3 is similar to itself. It is easily seen that only the first cluster contains relevant information, and thus the other two clusters redundant.

Acknowledgements

Financial support from BOF (Special Research Fund of Ghent University) is gratefully acknowledged.
I am grateful to Bart Wyns for his useful comments.

References

[1] J.C. Bezdeck, 1984, *FCM: fuzzy c-means algorithm*, Computers and Geoscience, **10**, pp. 191-203.
[2] R. C. Dubes and A. K. Jain, 1988, *Algorithms for clustering data*, Prentice Hall, 1988.
[3] X.Z. Fern and W. Lin, 2008, *Cluster ensemble selection*, Statistical Analysis and Data Mining, **1**, pp. 128-141.
[4] P. Hore, L.O. Hall and D.B. Goldgof, 2009, *A scalable framework for cluster ensembles*, Pattern Recognition, **42**, pp. 676-688.
[5] X. Hu and I. Yoo, 2004, *Cluster ensemble and its applications in gene expression analysis*, Proceedings of the second conference on Asia-Pacific bioinformatics, **29**, pp. 297-302.
[6] A.K. Jain, M. Murty and P. Flynn, 1999, *Data clustering: a review*, ACM Computing Surveys, **3**, pp. 264-323.
[7] A.K. Jain, 2010, *Data clustering: 50 years beyond K-means*, Pattern Recognition Letters, **31**, pp. 651-666.
[8] D. Jiang, C. Tang and A. Zhang, 2004, *Cluster analysis for gene expression data: a survey*, IEEE Transactions on Knowledge and Data Engineering, **16**, pp. 1370-1386.
[9] A. Strehl and J. Ghosh, 2002, *Cluster ensembles - a knowledge reuse framework for combining multiple partitions*, Journal of Machine Learning, **3**, pp. 583-617.
[10] X. Rui and D. Wunsch, 2005, *Survey of clustering algorithms*, IEEE Transactions on Neural Networks, **16**, pp. 645-678.
[11] K. Thangavel and N.K. Visalakshi, 2009, *Ensemble based distributed k-harmonic means clustering*, International Journal of Recent Trends in Engineering, **2**, pp. 125-129.
[12] G. F. Tzortzis and A. C. Likas, 2009, *The global kernel k-Means algorithm for clustering in feature space*, IEEE Transactions on Neural Networks, **20**, pp. 1181-1194.
[13] J.C. Venter *et al.*, 2001, *The sequence of the human genome*, Science, **291**, pp. 1304-1351.