Overview of overlapping partitional clustering methods

Chiheb-Eddine Ben N'Cir, Guillaume Cleuziou and Nadia Essoussi

Abstract Identifying non-disjoint clusters is an important issue in clustering referred to as Overlapping Clustering. While traditional clustering methods ignore the possibility that an observation can be assigned to several groups and lead to k exhaustive and exclusive clusters representing the data, Overlapping Clustering methods offer a richer model for fitting existing structures in several applications requiring a non-disjoint partitioning. In fact, the issue of overlapping clustering has been studied since the last four decades leading to several methods in the literature adopting many usual approaches such as hierarchical, generative, graphical and k-means based approach. We review in this paper the fundamental concepts of overlapping clustering while we survey the widely known overlapping partitional clustering algorithms and the existing techniques to evaluate the quality of non-disjoint partitioning. Furthermore, a comparative theoretical and experimental study of used techniques to model overlaps is given over different multi-labeled benchmarks.

1 Introduction

Clustering, also referred to as cluster analysis or learning, has become an important technique in data mining and pattern recognition used either to detect hidden structures or to summarize the observed data. Usually, a clear distinction is made

Chiheb-Eddine Ben N'Cir
LARODEC, ISG Tunis, University of Tunis, 41 Avenue de la Liberté, Cité Bouchoucha, Le Bardo-2000- Tunisia, e-mail: chiheb.benncir@isg.rnu.tn

Guillaume Cleuziou
LIFO, Université d'Orléans, EA 4022, Orléans, France e-mail: guillaume.cleuziou@univ-orleans.fr

Nadia Essoussi
LARODEC, ISG Tunis, University of Tunis, 41 Avenue de la Liberté, Cité Bouchoucha, Le Bardo-2000- Tunisia, e-mail: nadia.essoussi@isg.rnu.tn

between learning problems that are supervised, also referred to as classification, and those that are unsupervised, referred to as clustering. The first deals with only labeled data while the latter deals with only unlabeled data [Richard Duda, 2001]. In practice, given a description of N data over d variables, clustering aims to find k groups based on a measure of similarity such that similarities between data in the same group are high while similarities between data in different groups are low.

During the last four decades, many researches have been focused in designing clustering methods resulting in many methods that are proposed in the literature which are based on different approaches. Partitional clustering [Halkidi et al., 2001], also referred as partitioning relocation clustering [Berkhin, 2006], constitutes an important approach that several clustering methods are based on. Partitional clustering attempts to directly decompose the data set into a set of disjoint clusters leading to an integer number of clusters that optimizes a given criterion function. The criterion function may emphasize a local or a global structure of the data and its optimization is an *iterative* relocation procedure. Such type of clustering methods considers that clusters are disjoint and does not support intersections. However, for many applications of clustering, it would be recommended to tolerate overlaps between clusters to better fit hidden structures in the observed data. This research's issue is referred to as *overlapping clustering*.

Overlapping clustering has been studied through various approaches during the last half-century. In this paper, we first give a classification of existing methods able to produce a non-disjoint partitioning of data based on their conceptual approach. Then, we review overlapping clustering methods which extends or generalizes k-means and k-medoid for overlapping clustering. We use the concept "overlapping partitional clustering methods" to refer to this kind of methods which aims to build a recovery of a data set containing N objects into a set of k covers or clusters, so to minimize an objective criterion. The sum of the cardinality of the clusters are equal or superior to N leading to k non-exclusive clusters.

The remaining sections are organized as in the following: Section 2 gives a classification of existing overlapping methods based on the used approach to build non-disjoint partitioning. Then Section 3 reviews the existing overlapping partitional clustering methods while Section 4 reviews the existing techniques to asses the quality of the resulting non-exclusive partitionings. Finally, Section 5 gives an experimental evaluation of overlapping partitioning clustering methods on different multi-labeled benchmarks.

2 Classification of exiting approaches for overlapping clustering

Traditional learning methods ignore the possibility that an observation can belong to more than one cluster and lead to k exhaustive and exclusive clusters representing the data. Although this approach has been successfully applied in unsupervised learning, there are many situations in which a richer model is needed for representing the data. For example, in social network analysis, community extraction

algorithms need to detect overlapping clusters where an actor can belong to multiple communities [Tang and Liu, 2009, Wang et al., 2010, Fellows et al., 2011]. In video classification, overlapping clustering is a necessary requirement where videos have potentially multiple genres [Snoek et al., 2006]. In emotion detection, overlapping clustering methods need to detect different emotions for a specific piece of music [Wieczorkowska et al., 2006]. In text clustering, learning methods should be able to group document, which discuss more than one topic, into several groups [Gil-García and Pons-Porrata, 2010, Pérez-Suárez et al., 2013b], etc. The corresponding research domain has been referred to as *overlapping clustering* and has been studied through various approaches.

Basically, the existing overlapping clustering methods are extensions from usual clustering models such as hierarchical, generative, graph-based or partitional models. Thereby, we propose a classification of existing methods based on the conceptual approach to build non-disjoint partitioning of data. Figure 1 shows a classification tree of these methods where the depth of the tree represents the progression in time and the width of the tree represents the different categories and subcategories. We detail in the following the main characteristics of each category.

The overlapping variants of hierarchies aim to reduce the discrepancy between the original dissimilarities over the considered dataset and the ones induced by the hierarchical structure. Although the flexibility in visualization offered by hierarchical methods, they still too restrictive in overlaps while they not study all the possible combinations of clusters for each observation. Examples of these methods are the pyramids [Diday, 1984] and more generally the weak-hierarchies [Bertrand and Janowitz, 2003]. Concisely, these structures are either too restrictive on the overlaps, as for pyramids, or hard to acquire and visualize as for weak-hierarchies.

Overlapping methods based on graph theory are mostly used in the context of community detection in complex networks [Gregory, 2007, Zhang et al., 2007a, Baumes et al., 2005, Magdon-Ismail and Purnell, 2011, Davis and Carley, 2008, Goldberg et al., 2010, Wang and Fleury, 2011, Gregory, 2008, Fellows et al., 2011]. In this research area, a network is represented as an undirected or directed graph, depending on the specificity of the problem, where vertices are the studied observations and edges are links between the observations. All these graph-based methods use a greedy heuristic for covering the similarity graph. The difference between them consists on the criterion used for ordering and selecting the sub-graphs. The main shortcoming of these methods is the computational complexity which is usually exponential and could be reduced to $O(N^2)$ as the case for OClustR (Overlapping Clustering based on Relevance) [Pérez-Suárez et al., 2013a].

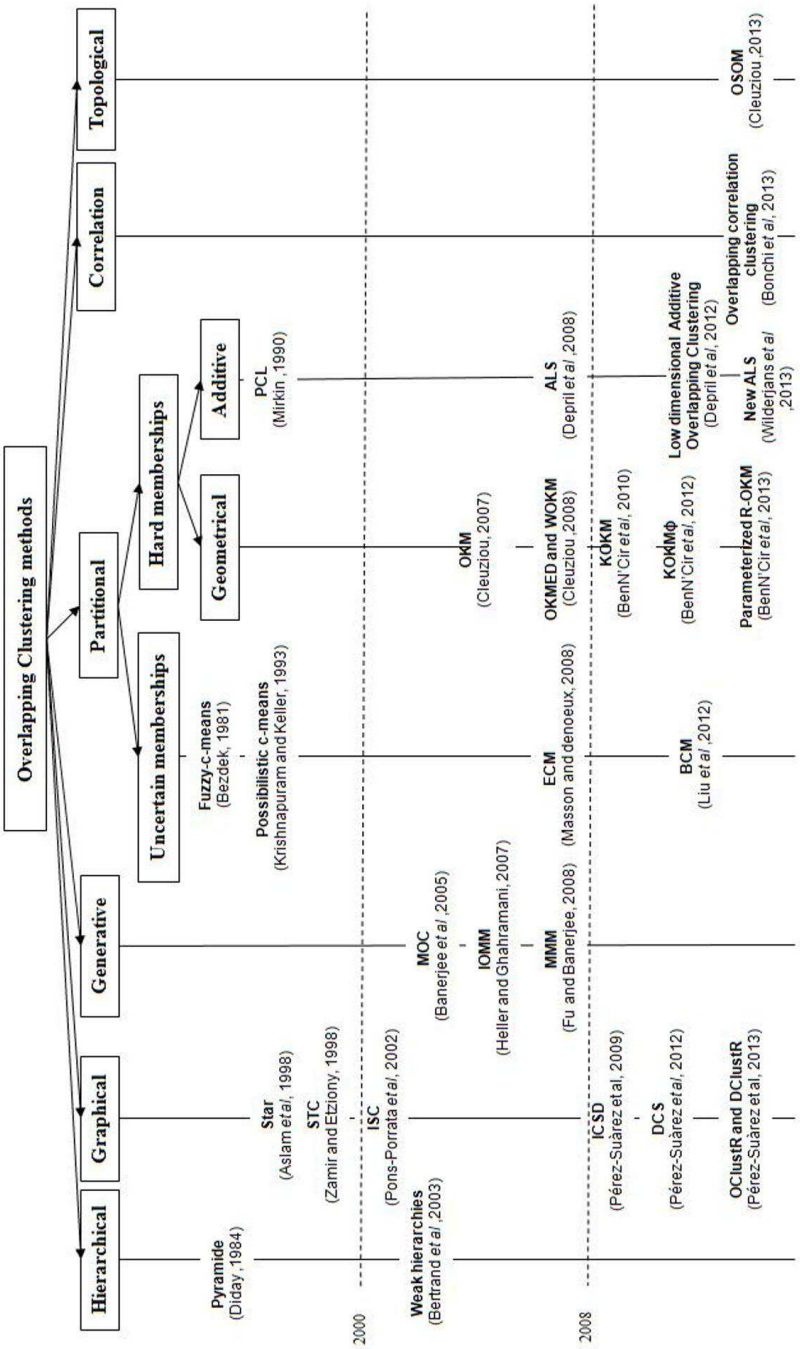


Fig. 1 Classification of overlapping clustering methods based on their conceptual approach to build non-disjoint partitioning of data.

Overlapping clustering methods using generative mixture models have been proposed [Banerjee et al., 2005, Heller and Ghahramani, 2007, Fu and Banerjee, 2008] as extensions of the EM algorithm [Dempster et al., 1977]. These models are supported by biological processes; they hypothesize that each data is the result of a mixture of distributions : the mixture can be additive [Banerjee et al., 2005] or multiplicative [Heller and Ghahramani, 2007, Fu and Banerjee, 2008] and the probabilistic framework makes possible to use not only gaussian components but any exponential family distributions. On the other hand, generative models are not parameterizable and do not allow the user to control the requirements of the overlaps.

Other recent methods for overlapping clustering extend other approaches to address the problem of overlapping clustering. For example, an extension of correlation clustering [Bonchi et al., 2011, Bonchi et al., 2013] and topological maps [Cleuziou, 2013] have been recently proposed. Overlapping correlation clustering has been defined as an optimization problem that extends the framework of correlation clustering to allow overlaps by relaxing the function which measures the similarity of assigned set of labels, instead of one single label, for each data object. For topological maps, an extension of the Self-Organizing-Maps (SOM) has been proposed by allowing for each data to be assigned to one or several neurons on the grid by searching for a subset of neurons winners rather than a single neuron. The main advantage of both correlation and topological methods consists of their ability to learn the right number of overlapping clusters.

Despite the use of all these approaches to build non-disjoint partitioning of data, the Partitional approach remains the most commonly used while several methods are based on. This category of methods consists either in modifying the clusters resulting from a standard method into overlapping clusters or in proposing new objective criteria to model overlaps. This survey's emphasis is on overlapping partitional clustering methods, specifically those extending and generalizing k-means and K-medoid methods[MacQueen, 1967, Jain, 2010]. In this way, we present in the next section a description of these methods.

3 Overlapping partitional clustering methods

Several works have been focused on partitional clustering to build overlapping clusters leading to two main categories of methods: the category of *uncertain memberships* and the category of *hard memberships*. We denote by uncertain memberships the solutions which model clusters' memberships for each data object as uncertainty function using fuzzy, possibilistic or evidential frameworks. The uncertainty function measures the degree of belonging of each data to the underlying group. However, we denote by *hard memberships* the solutions which lead to *hard* and *overlapping* partitioning by considering a binary function to model clusters' memberships.

3.1 Uncertain memberships based-methods

Uncertain memberships based-methods consist either in extending results of uncertain methods into overlapping clusters, typically the extension of fuzzy-*c*-means (FCM) [Lingras and West, 2004, Zhang et al., 2007a] and possiblistic c-means (PCM) [Krishnapuram and Keller, 1993], or in proposing new objective criterion that takes into account the possibility of overlaps between clusters; the Evidential c-means (ECM) introduced by [Masson and Denoeux, 2008] and the Belief c-means (BCM) proposed by [Liu et al., 2012] are two distinctive examples of such criteria where their optimization processes lead to generate overlapping clusters. All uncertain methods need a post-processing treatment to generate the final overlapping clusters.

We detail in the following the principal uncertain clustering methods which are able to produce non-disjoint partitioning. We consider for all the detailed methods a set of observations $X = \{x_i\}_{i=1}^{N}$ with $x_i \in \mathbb{R}^d$ and N the number of observations where the aim, of each method, is to find a non-disjoint partitioning matrix $\Pi = \{\pi_c\}_{c=1}^{k}$ into k clusters and a set $C = \{m_c\}_{c=1}^{k}$ of k clusters' representatives minimizing an objective criterion.

3.1.1 Fuzzy c-means (FCM) and Possiblistic c-means (PCM)

The FCM [Bezdek, 1981] identifies clusters as fuzzy sets where the objective function J_{FCM} allows that an observation belongs to many clusters with a coefficient indicating membership degrees to all clusters in the [0,1] interval (0 stands for no membership and 1 for total membership). FCM is based on the minimization of the following function:

$$J_{FCM}(\Pi, C) = \sum_{c=1}^{k} \sum_{i=1}^{N} \pi_{ic}^{\beta} . ||x_i - m_c||^2, \tag{1}$$

where Π is the fuzziness membership matrix that indicates the coefficient of closeness of an object to every cluster under the constraints:

$$\begin{gathered} \pi_{ic} \in [0..1] \ \forall i, \quad \forall c \\ \sum_{c=1}^{k} (\pi_{ic}) = 1, \forall i \\ \beta > 1. \end{gathered} \tag{2}$$

The parameter β controls the fuzziness of the memberships: for high values of β the algorithm tends to set all the memberships equals while for β tending to one it has the behavior of k-means algorithm with crisp memberships. The minimization of Equation 1 is done iteratively using an alternating least square optimization of

the two parameters Π and C. The optimal fuzziness membership matrix Π and the optimal clusters' representatives C are computed in each step as the following.

$$\pi_{ic}^{*} = \frac{1}{\sum_{l=1}^{k} \left(\frac{||x_i - m_c||^2}{||x_i - m_l||^2} \right)^{\left(\frac{1}{\beta-1}\right)}} \tag{3}$$

$$m_c^{*} = \frac{\sum_{i=1}^{N} \pi_{ic}^{\beta} x_i}{\sum_{i=1}^{N} \pi_{ic}^{\beta}} \tag{4}$$

The extension of FCM to overlapping clustering can be done by fixing a threshold where all observations having memberships' degrees that exceed this threshold are assigned to the respective clusters. An example of this transformation is shown in Figure 2 where the obtained clusters' memberships are non-disjoints. We note that in some cases, when the fixed threshold is somewhat large, observations having all the memberships lower than this threshold will not be assigned to any cluster. Obtained clusters are usually much sensitive to value of the threshold.

	Fuzzy Clustering		
observation	Cluster1	Cluster2	Cluster3
1	0.05	0.90	0.05
2	0.50	0.20	0.30
3	0.30	0.40	0.30
4	0.70	0.20	0.10
5	0.20	0.40	0.40
6	0.00	0.20	0.80

Threshold = 0.3

	Overlapping Clustering		
observation	Cluster1	Cluster2	Cluster3
1	0	1	0
2	1	0	1
3	1	1	1
4	1	0	0
5	0	1	1
6	0	0	1

Fig. 2 Extension of the results of fuzzy clustering to obtain overlapping clustering using a threshold value equals to 0.3

In the same way that FCM, the PCM [Krishnapuram and Keller, 1993] method is proposed to relax the constrained condition of the fuzzy partition ($\sum_{c=1}^{k}(\pi_{ic}) = 1$) in order to obtain a possiblistic type of memberships matrix. The objective function of PCM[1] is described by:

$$J_{PCM}(\Pi, C) = \sum_{c=1}^{k}\sum_{i=1}^{N} \pi_{ic}^{\beta} ||x_i - m_c||^2, \tag{5}$$

under the constraints:

$$\begin{aligned} &\pi_{ic} \in [0..1] \ \forall i, \quad \forall c \\ &\sum_{c=1}^{k}(\pi_{ic}) \quad \in [0,k] \ \forall \ i \\ &\beta > 1. \end{aligned} \tag{6}$$

Similar to FCM, the objective function J_{PCM} is minimized iteratively where memberships and clusters' representatives are updated as follows:

$$\pi_{ic}^{*} = \frac{1}{1 + (||x_i - m_c||^2)^{1/(\beta-1)}}, \tag{7}$$

$$m_c^{*} = \frac{\sum_{i=1}^{N} \pi_{ic}^{\beta} x_i}{\sum_{i=1}^{N} \pi_{ic}^{\beta}}. \tag{8}$$

Each observation can belong to several clusters with a possiblistic membership. The possiblistic memberships can be extended to overlapping ones by setting memberships with higher values to 1 and the other are replaced with 0.

3.1.2 Evidential c-means (ECM) and Belief c-means (BCM)

ECM [Masson and Denoeux, 2008] is based on the concept of credal partition which is a general extension of fuzzy and possiblistic partitioning. As opposed to FCM and PCM, the ECM method evaluates all the possible combinations of clusters , denoted A_j, from the set of single clusters $\Omega = \{\omega_1, \ldots, \omega_k\}$ by allocating a mass of belief π_i within each possible combination.

[1] The original objective function of PCM takes into account the identification of outliers, whereas we give in this paper a short structure of the objective function to facilitate the comparison of PCM with the other described methods

Let the credal partition matrix $\Pi = (\pi_1, \ldots, \pi_N) \in R^{N \times 2^k}$ and the matrix $C = (m_1, \ldots, m_k)$ of clusters' representatives, ECM[2] is based on the minimization of the following objective function:

$$J_{ECM}(\Pi, C) = \sum_{i=1}^{N} \sum_{j/A_j \subseteq \Omega} c_j^{\alpha} \pi_{ij}^{\beta} \left\| x_i - \bar{m}_j \right\|^2 \tag{9}$$

under the constraint:

$$\sum_{j/A_j \subseteq \Omega} \pi_{ij} = 1 \; \forall i \in \{1, .., N\}, \tag{10}$$

where π_{ij} denotes the mass of belief for associating the observation x_i to the specific set A_j which can be either a single cluster or a combination of single clusters, c_j denotes the cardinality $|A_j|$ of each set which aims at penalizing the combination of clusters with high cardinality, α and β are two parameters used respectively to control the penalization term c_j and the fuzziness degree π_{ij} and $\left\| x_i - \bar{m}_j \right\|^2$ is the distance between x_i and the combination of clusters' representatives $\overline{m_j}$ of the focal set A_j defined by:

$$\overline{m_j} = \frac{\sum_{j/A_j \subseteq \Omega} m_j}{c_j}. \tag{11}$$

To minimize the objective function of ECM, an alternating optimization scheme is designed as in FCM where the update of the mass of belief π_{ij} and the clusters' representatives C_k is computed as described in Equations 12 and 13 .

$$\pi_{ij}^{*} = \frac{c_j^{-\alpha/(\beta-1)} \left\| x_i - \bar{m}_j \right\|^{-2/(\beta-1)}}{\sum_{A_k} c_k^{-\alpha/(\beta-1)} \left\| x_i - \bar{m}_k \right\|^{-2/(\beta-1)}} \; \forall i \in \{1, .., N\} \; \forall j / A_j \subseteq \Omega, \tag{12}$$

$$C_{k \times d}^{*} = H_{k \times k}^{-1} B_{k \times d}, \tag{13}$$

where the elements B_{lq} of the matrix $B_{k \times d}$ for $l \in \{1, .., k\}$, $q \in \{1, .., d\}$ and the elements H_{lh} of the matrix $H_{k \times k}$ for $l, h \in \{1, .., k\}$ are defined respectively by:

$$B_{lq} = \sum_{i=1}^{N} x_{iq} \sum_{j/\omega_l \in A_j} c_j^{\alpha-1} \pi_{ij}^{\beta} \qquad H_{lh} = \sum_{i=1}^{N} \sum_{j/\omega_h, \omega_l \subseteq A_j} c_j^{\alpha-2} \pi_{ij}^{\beta}. \tag{14}$$

[2] The original objective function of ECM takes into account the identification of outliers by considering $\pi_{i\emptyset}$ a mass of belief to belong to any cluster, whereas we consider in this paper that all combinations of clusters are tolerated except the empty set ($A_j \neq \emptyset$) in order to facilitate the comparison of ECM with the other described methods

Similar to ECM, a more recent method, referred to as Belief c-means (BCM) [Liu et al., 2012], is also developed within the framework of belief function which is an extension of ECM that improves the quality of the credal partitions by assigning only relatively distant observations to the combination of clusters. This method evaluates the distance inter-prototypes before building the mass of believe π_{ij} related to each observation. The objective function optimized by BCM is described by:

$$J_{BCM}(\Pi,C) = \sum_{i=1}^{N} \sum_{j/A_j \subseteq \Omega, |A_j|=1} \pi_{ij}^{\beta} \left\|x_i - m_j\right\|^2 + \sum_{i=1}^{N} \sum_{j/A_j \subseteq \Omega, |A_j|>1} c_j^{\alpha} \pi_{ij}^{\beta} \widehat{d_{ij}}^2, \quad (15)$$

where $\widehat{d_{ij}}^2$ evaluates the distance x_i with respect to inter-distances of clusters' prototypes to which x_i belongs to:

$$\widehat{d_{ij}}^2 = \frac{\sum_{k \in A_j} \|x_i - m_k\|^2 + \sum_{l,p \in A_j} \left\|m_l - m_p\right\|^2}{|A_j| + \gamma C_{|A_j|}^2}, \quad (16)$$

with γ the weighting factor of the distances among the clusters' prototypes and $C_{|A_j|}^2 = \frac{|A_j|!}{2!(|A_j|-2)!}$ the number of combinations of $|A_j|$ taken 2 at a time.

In fact, both ECM and BCM lead to credal partitioning of N data into 2^k possible combinations of k clusters. An example of credal partitioning is reported in Figure 3. We note that both possiblistic and fuzzy partitions can be recovered from the credal partition. A possiblistic partition can be obtained by computing from each m_i the plausibilities (possibilities) of the different clusters and a fuzzy partition can be obtained by calculating the pignistic probabilities of the different clusters from each m_i. The extension of both ECM and BCM to overlapping clustering, called hard credal partition by the authors, can be done by assigning each object x_i to the set of clusters A_j with the highest mass. Overlapping observations are those having highest mass for $|A_j| >= 2$. The process that leads to hard credal partitioning is called "Upper" [Masson and Denoeux, 2008] approximation.

3.2 Hard memberships based-methods

The category of hard memberships based methods generalizes the strict k-means to look for optimal overlapping clusters. As opposed to fuzzy, possiblistic, and evidential clustering, these methods produce hard overlapping clusters and do not need any post processing treatment. Two kind of hard memberships based-methods have been proposed: *additive* and *geometrical* based methods. We denote by *additive* the methods which hypothesize that overlaps result in the addition of the representatives of the related clusters. These methods group observations into overlapping clusters

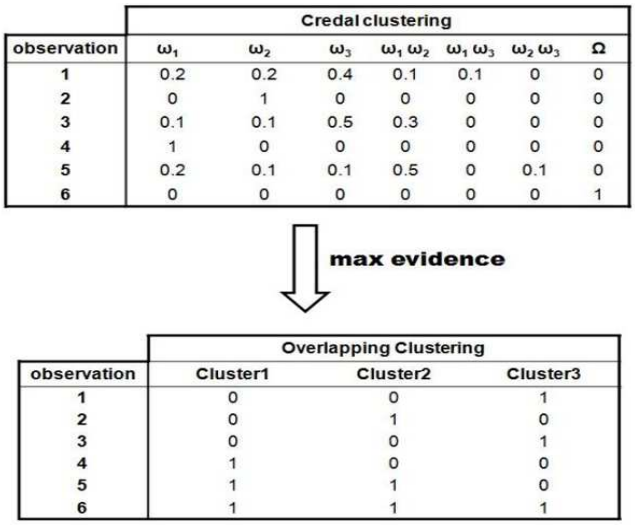


Fig. 3 Example of the extension of credal partition containing 6 observations under 3 clusters: observations are assigned to the combination of clusters having the max mass of evidence.

while minimizing the sum of distances between each observation and the *sum* of clusters' representatives to which the observation belongs to. In contrast, we denote by *geometrical* based-methods those formalizing overlaps as a barycenter on the related cluster representatives. This category of methods is based on a geometrical reasoning in the data space and groups observations into overlapping clusters while minimizing the sum of distances between each observation and the *average*, instead of the sum, of clusters' representatives to which the observation belongs to.

3.2.1 Additive methods

Examples of additive methods are Principal Cluster Analysis (PCL) [Mirkin, 1987, Mirkin, 1990], the Alternating Least Square algorithms (ALS) [Depril et al., 2008], the Lowdimensional Additive Overlapping Clustering [Depril et al., 2012] and the Bi-clustering ALS [Wilderjans et al., 2013].

- **PCL**
  PCL [Mirkin, 1990] introduces the possibility that an observation belongs to more than one cluster based on the Additive model by considering variable values of an observation equals to the sum of the clusters' representatives to which the observation belongs to. Given a dataset X, a model matrix $M = \Pi C$ is looked for to optimally approximate X. The matrix M can be estimated by minimizing the least squares loss function:

$$J_{PCL}(\Pi, C) = \| X - \Pi C \|_F^2 = \sum_{x_i \in X} \| x_i - \sum_{c \in \Pi_i} m_c \|^2, \quad (17)$$

where $\| \cdot \|_F^2$ is the Frobenius norm of a matrix and m_c is the representative of cluster c.

To minimize the objective criterion (17), PCL proceeds by building clusters one by one from a data set until achieving the expected number of clusters k. PCL builds the memberships of each cluster c independently from the memberships of the other clusters by minimizing the following criterion for each cluster c:

$$J_{PCL}^c = \sum_{x_i \in X} \pi_{ic} \| x_i - m_c \|^2 . \quad (18)$$

The minimization process starts with an empty cluster (i.e., with $\pi_{ic} = 0, \forall x_i \in X$) and sequentially add observations to it in a greedy way leading to the smallest value of criterion (18). For every observation that is considered for joining the cluster, a new representative m_c and a new value of J_{PCL}^c have to be calculated. The process is continued until there is no further decrease of J_{PCL}^c. The computation of clusters' representatives is also done locally for each cluster independently from the other clusters by:

$$m_c^* = \frac{\sum_{x_i \in X} \pi_{ic} x_i}{\sum_{x_i \in X} \pi_{ic}}. \quad (19)$$

The main characteristic of this method consists of its high computational complexity evaluated by $O(2^{NK})$ which makes the method under-used in real life applications of overlapping clustering.

- **ALS**

ALS [Depril et al., 2008] is based on the same objective criterion of PCL described in Equation 17. However, ALS proposes two other algorithms for minimizing this objective criterion referred to as ALS_{lf1} and ALS_{lf2}. For both algorithms, the minimization of the objective criterion starts from an initial binary membership matrix Π_0. This membership matrix can be initialized using a Bernoulli distribution with parameter $\tau = 0.5$ or by computing the conditionally optimal memberships upon k randomly drawn representatives from the initial data. Then, ALS estimates the conditionally optimal representatives C upon Π; subsequently it estimates the conditionally optimal memberships Π upon C, and this process will be repeated until convergence.

The estimation of optimal memberships are computed separably for each observation x_i by enumerating all possible binary configurations that lead to decrease the objective criterion given the conditionally optimal representative and the conditionally optimal memberships for the other observations. The algorithm repeats this procedure for the next observation and so on. After a pass through all observations, the new value of the objective criterion computed with new mem-

berships Π_1 is compared to the one computed with old memberships Π_0. The process stops when there is no further decrease in the objective criterion. We note that ALS$_{lf1}$ differs from ALS$_{lf2}$ in that for each memberships update Π_i the conditionally optimal representatives C are recalculated immediately, whereas in ALS$_{lf2}$ the representatives are only updated at the end of the membership updating step.
In the other side, the optimal representatives are updated equivalently for ALS$_{lf1}$ and ALS$_{lf2}$ based on the memberships matrix Π as follows:

$$C^* = (\Pi'\Pi)^{-1}\Pi'X. \tag{20}$$

In fact, we notice that ALS with its two variants explores all the possible 2^k combinations of clusters and takes the optimal one that leads to decrease the objective criterion. This property makes this methods highly time consuming when the number of clusters becomes large.

- **Low dimensional Additive Overlapping Clustering**
The Low dimensional Additive Overlapping Clustering [Depril et al., 2012] extends ALS method by establishing an overlapping clustering of the observations and a dimensional reduction of the variables (or dimensions) simultaneously. This method is designed in order to perform relevant non-disjoint partitioning when data contains a high number of dimensions. Given a set of observations X described over d variables, the aim of this method is to find a recovery Π of k overlapping clusters and a matrix $\widetilde{C}$ of clusters' representatives described over $\widetilde{d} < d$. The low dimensional additive Overlapping Clustering is based on the same objective criterion used for ALS and PCL. The process of optimizing this objective criterion uses an alternating optimization procedure similar to that of ALS, except that optimal reduced clusters' representatives are computed by:

$$\widetilde{C}^* = (\Pi'\Pi)^{-1}\Pi'TT'X, \tag{21}$$

where columns of matrix T represent the $\widetilde{d}$ orthogonal eigenvectors of the product of matrixes $ZXX'Z$ with $Z = \Pi(\Pi'\Pi)^{-1}\Pi'$ denotes the orthogonal projector operator.

3.2.2 Geometrical methods

Examples of geometrical methods are the Overlapping k-means (OKM) [Cleuziou, 2008], Overlapping k-Medoid (OKMED) [Cleuziou, 2009], Weighted Overlapping k-means (WOKM) [Cleuziou, 2009], Kernel Overlapping K-means (KOKM) [BenN'Cir et al., 2010, BenN'Cir and Essoussi, 2012] and Parameterized R-OKM [BenN'Cir et al., 2013]. We also note that ECM and BCM, described with uncertain memberships methods, can be categorized with geometrical methods while they are based on a geometrical reasoning to model the different combinations of clusters.

- **OKM**
  The OKM [Cleuziou, 2008] method is an extension of the k-means algorithm which allows observations to belong to one or different clusters. Given a set of N observations $X = \{x_i\}_{i=1}^{N}$ with $x_i \in \mathbb{R}^d$, OKM aims to find a recovery $\Pi = \{\pi_c\}_{c=1}^{k}$ of k overlapping clusters such that the following objective function is minimized:

$$J_{OKM}(\Pi, C) = \sum_{i=1}^{N} \|x_i - \overline{(x_i)}\|^2. \tag{22}$$

  This objective function minimizes the sum of squared Euclidean distances between each observation x_i and its representatives $\overline{(x_i)}$ for all $x_i \in X$. The representative $\overline{(x_i)}$ is defined as the barycenter of clusters' representatives to which the observation x_i belongs to:

$$\overline{(x_i)} = \sum_{c \in \Pi_i} \frac{m_c}{|\Pi_i|}. \tag{23}$$

  where Π_i is the set of clusters to which x_i belongs to and m_c is the representative of cluster c. The minimization of the objective function is performed by alternating two principal steps: computation of clusters' representatives (C) and the assignment of observations to one or several clusters (Π). The update of representatives is performed locally for each cluster as described in Equation 24.

$$m_c^* = \frac{\sum_{x_i \in \pi_c} \frac{1}{|\Pi_i|^2} . \tilde{x}_i^c}{\sum_{x_i \in \pi_c} \frac{1}{|\Pi_i|^2}}, \tag{24}$$

  where $\tilde{x}_i^c = |\Pi_i| . x_i - (|\Pi_i| - 1) . \overline{(x_i)_{\Pi \setminus c}}$ and π_c denotes the set of observations assigned to cluster c. For the multiple assignment step, OKM uses an heuristic to explore part of the combinatorial set of possible assignments. The heuristic consists, for each observation, in sorting clusters from closest to the farthest, then assigning the observation in the order defined while assignment minimizes the distance between the observation and its representative.

- **Parameterized R-OKM**
  The Parameterized R-OKM offers the possibility to regulate the overlaps using a parameter α. As well as α becomes large ($\alpha \to +\infty$), Parameterized R-OKM builds clusters with more reduced overlaps. However, overlaps become more large as well as $\alpha \to 0$. When $\alpha = 0$, Parameterized R-OKM coincides exactly with OKM. In fact, Parameterized R-OKM restricts the assignments of a data point x_i to multiple clusters according to the cardinality of the set of assignments $|\Pi_i|$. The Parameterized R-OKM is based on the minimization of the following objective criterion:

$$J_{R-OKM}(\Pi,C) = \sum_{i=1}^{N} |\Pi_i|^{\alpha}.||x_i - (\overline{x_i})||^2 \tag{25}$$

where $\alpha \geq 0$ is fixed by the user. We note that $|\Pi_i|^{\alpha}$ has the same role of C_j^{α} used within ECM and BCM methods which consists on penalizing or favoring overlaps. For the minimization of the objective criterion, Parameterized R-OKM uses the same minimization steps used for OKM as for the assignment and for the update of clusters' representatives. The later can be updated for each cluster as follows:

$$m_c^* = \frac{\sum_{x_i \in \pi_c} \frac{1}{|\Pi_i|^{2-\alpha}} . \tilde{x_i}^c}{\sum_{x_i \in \pi_c} \frac{1}{|\Pi_i|^{2-\alpha}}}. \tag{26}$$

- **Kernel Overlapping k-means**
  While most of geometrical overlapping methods build non-disjoint partitioning of data with linear separations between clusters, KOKM hypothesizes that data structuring in real life applications is usually complex; thus requiring nonlinear separations to better fit the existing structures in data. In order to perform nonlinear separations between clusters, KOKM introduces the use of kernel methods for overlapping clustering. Two variants are proposed: the first [BenN'Cir et al., 2010] proposes a kernelization of the Euclidean metric used in OKM using the kernel induced distance measure while the second [BenN'Cir and Essoussi, 2012], referred as KOKMϕ, proposes to perform all clustering steps in a high dimensional space where data are implicitly mapped.
  Given a set of observations $X = \{x_i\}_{i=1}^N$ and given an implicit nonlinear mapping function $\phi : X \longrightarrow F$ which maps the input space X to a high dimensional feature space F, the objective functions minimized by both variants are respectively described by:

$$J_{KOKM}(\Pi,C) = \sum_{i=1}^{N} \|\phi(x_i) - \phi(\overline{x_i})\|^2, \tag{27}$$

$$J_{KOKM\phi}(\Pi,C) = \sum_{i=1}^{N} \|\phi(x_i) - \overline{\phi(x_i)}\|^2. \tag{28}$$

  The first variant computes representatives $\overline{x_i}$ and clusters' representatives m_c in the original space and only distances between observations are performed in the mapping space. The optimization steps are similar to OKM except that distances are computed in feature space as described in Equation 29.

$$\begin{aligned} \| \phi(x_i) - \phi(x_j) \|^2 &= (\phi(x_i) - \phi(x_j))((\phi(x_i) - \phi(x_j)) \\ &= \phi(x_i)\phi(x_i) - 2\phi(x_i)\phi(x_j) + \phi(x_j)\phi(x_j) \\ &= K_{ii} - 2K_{ij} + K_{jj}. \end{aligned} \tag{29}$$

where K_{ij} is the dot product of mapped data in the feature space which can be computed without using ϕ.
Conversely, the second variant performs all the learning process in the feature space F. The representatives $\overline{\phi(x_i)}$ are computed in feature space as follows:

$$\overline{\phi(x_i)} = \frac{\sum_{c=1}^{k} \pi_{ic}.m_c^{\phi}}{\sum_{c=1}^{k} \pi_{ic}}, \tag{30}$$

where $\pi_{ic} \in \{0,1\}$ is a binary variable that indicates membership of observation x_i to cluster c and m_c^{ϕ} is the representative of cluster c in the feature space. The representative of each cluster in the feature space is defined by the medoid that minimizes all distances over all observations included in this cluster:

$$m_c^* = \underset{x_i \in \pi_c}{\arg\min}(x_i) \frac{\sum_{x_j \in \pi_c, x_j \neq x_i} |\Pi_j|[K_{ii} - 2K_{ij} + K_{jj}]}{|\pi_c|. \sum_{x_j \in \pi_c, x_j \neq x_i} |\Pi_j|}. \tag{31}$$

- **OKMED**
  OKMED extends the method Partitioning Around Medoid (PAM) for overlapping clustering. It consists in aggregating the data around representatives of the clusters denoted as medoids which are chosen among the data themselves. The objective criterion of OKMED is based on the optimization of the following objective criterion:

$$J_{OKMED}(\Pi, C) = \sum_{x_i \in X} \|x_i - \overline{(\chi_i)}\|^2. \tag{32}$$

where $\overline{(\chi_i)}$ is defined as the data from X that minimizes the sum of the dissimilarities with all the medoids of the clusters where x_i belongs to:

$$\overline{(\chi_i)} = \underset{x_j \in X}{\arg\min} \sum_{m_c \in A_i} \|x_j - m_c\|^2. \tag{33}$$

The optimization of the objective function of OKMED is realized using an alternating optimization between two steps: assignment of each data to its nearest medoid and updating of the medoid for each cluster. The update of medoids consist in searching among the set of data belonging to the cluster, the one that minimizes the sum of the distances with any other data into the cluster. Formally, optimal medoid for each cluster is described by:

$$m_c^* = \underset{x_i \in \pi_c}{\arg\min}(x_i) \sum_{x_j \in \pi_c} \|x_j - \overline{(\chi_j)_{x_i}}\|^2, \tag{34}$$

where $\overline{(\chi_j)_{x_i}}$ denotes the representative of observation x_j computed by considering x_i the medoid of the cluster. The use of medoids as representatives of clusters makes OKMED more robust to outliers and offers the possibility to use any metric since it only requires a proximity matrix over the data.

- **WOKM**
  The WOKM [Cleuziou, 2009] method is a generalization of both OKM and Weighted k-means methods to detect overlapping clusters. The WOKM introduces a vector of local feature weighting λ_c, relative to each cluster c, which allows data to be assigned to a cluster as regards to a subset of attributes that are important for the cluster concerned. WOKM is based on the minimization of the following objective function:

$$J_{WOKM}(\Pi, C) = \sum_{i=1}^{N} \sum_{v=1}^{d} \gamma_{i,v}^{\beta} \|x_{i,v} - \overline{(x_{i,v})}\|^2, \tag{35}$$

where d the number of features and γ_i a vector of weights relative to the representative $\overline{(x_i)}$ which is defined for each feature v as the following:

$$\gamma_{i,v} = \frac{\sum_{c \in \Pi_i} \lambda_{c,v}}{|\Pi_i|}. \tag{36}$$

The representative $\overline{(x_{i,v})}$ is defined, for each feature v, as the weighted barycenter of clusters' representatives to which the observation x_i belongs to:

$$\overline{(x_{i,v})} = \frac{\sum_{c \in \Pi_i} \lambda_{c,v}^{\beta}.m_{c,v}}{\sum_{c \in \Pi_i} \lambda_{c,v}^{\beta}}. \tag{37}$$

The optimization of the objective criterion is performed by iterating three steps. The first step consists in assigning each data object to the nearest cluster while minimizing the local error $\sum_{v=1}^{d} \gamma_{i,v}^{\beta} \|x_{i,v} - \overline{(x_{i,v})}\|^2$. The second step consists in updating clusters' representatives using the following criterion:

$$m_{c,v}^* = \frac{\sum_{x_i \in \pi_c} \frac{\lambda_{i,v}^{\beta}}{|\Pi_i|^2}.\tilde{x}_i^c}{\sum_{x_i \in \pi_c} \frac{\lambda_{i,v}^{\beta}}{|\Pi_i|^2}}. \tag{38}$$

The third step concerns the update of the set of clusters weights $\{\lambda_c\}_{c=1}^{k}$ by using:

$$\lambda_{c,v} = \frac{(\sum_{x_i \in \pi_c} \|x_{i,v} - m_{c,v}\|^2)^{1/(1-\beta)}}{\sum_{u=1}^{d} (\sum_{x_i \in \pi_c} \|x_{i,u} - m_{c,u}\|^2)^{1/(1-\beta)}}. \quad (39)$$

The computational complexity of WOKM stills linear, similar to OKM, evaluated by $O(N.k.\lg k)$.

3.3 Summary of overlapping partitional methods

This section offers an overview of the main characteristics of overlapping partitional clustering methods presented in a comparative way. Table 1 summarizes these main characteristics. Our study is based on the following features of the methods: 1) model of overlaps in the objective criterion, 2) requirement of the method to use a post-assignment step to generate the final overlapping clusters, 3) type of clusters' representatives, 4) type of data supported by each method, 5) type of separations between clusters, 6) computational complexity, 7) ability to handle noise and outliers and 8) ability to regulate the sizes of overlaps.

Methods which integrate overlaps in their optimized criteria lead to two main categories, additive and geometrical, which differers in the assumptions and in the context of use. The adoption of additive or geometrical methods is motivated by the requirement of the application. Additive-based methods have been well applied in grouping patients into diseases. Each patient may suffer from more than one disease and therefore could be assigned to multiple syndrome clusters. Thus, the final symptom profile of a patient is the sum of the symptom profiles of all syndromes he is suffering from. However, this type of methods needs sometimes to prepare data to have zero mean to avoid false analysis. For example, if symptom variable represents the body temperature, then when a patient simultaneously suffers from two diseases, it is not realistic to assume that his body temperature equals to the sum of body temperatures as associated with two diseases.

Conversely, geometrical-based methods have been well applied to group music signals into different emotions and films into several genres. These methods consider that overlapping observations must appear in the extremity surface between overlapping clusters. For example, if a film belongs to action and horror genres, it should have some shared properties with these categories of films but it can neither be a full action film neither a full horror one. So, overleaping films belonging to action and horror categories may appear in the limit surface between full horror and full action films.

Table 1 The main characteristics of the overlapping partitional clustering methods.

Method	Overlap model	Post-assignment step	Representatives	Type of data	Separation between clusters	Complexity	Outliers identification	overlap regulation
FCM	-	threshold	centroid	numeric	linear	$O(N.k)$	no	yes
PCM	-	threshold	centroid	numeric	linear	$O(N.k)$	yes	yes
ECM	geometrical	max evidence	centoid	numeric	linear	$O(N.2^k)$	yes	yes
BCM	geometrical	max evidence	centoid	numeric	linear	$O(N.k.2^k)$	yes	yes
OKM	geometrical	-	centroid	numeric	linear	$O(N.k.\lg k)$	no	no
P. R-OKM	geometrical	-	centroid	numeric	linear	$O(N.k.\lg k)$	no	yes
KOKM	geometrical	-	centroid	any type	nonlinear	$O(N.k.\lg k)$	no	no
KOKMϕ	geometrical	-	medoid	any type	nonlinear	$O(N^2.k)$	no	no
OKMED	geometrical	-	medoid	any type	linear	$O(N^3.k)$	no	no
PCL	additive	-	centroid	numeric	linear	$O(2^{N.k})$	no	no
ALS$_{lf1}$	additive	-	centroid	numeric	linear	$O(N^3.2^k)$	yes	no
ALS$_{lf2}$	additive	-	centroid	numeric	linear	$O(N^2.2^k)$	yes	no

While all described overlapping partitional clustering methods offer a richer model to fit the existing structures in data, some parameters need to be estimated before performing the learning. All the described methods require to configure the number of clusters in prior which is not a trivial task in real life applications where the number of expected clusters is usually unknown. As a solution, one could use different model heuristics for determining the optimal number [Depril et al., 2012, Wilderjans et al., 2013]. For example, the user can test different clusterings with increasingly number of clusters and then, takes the clustering having the best balance between the minimization of the objective function and the number of clusters [Wilderjans et al., 2011]. Furthermore, all the described overlapping partitional methods need to initialize the clusters' representatives or the primary clusters memberships. Clusters' representatives can be set randomly from data themselves or can be determined using existing initialization methods [Celebi and Kingravi, 2012, Celebi et al., 2013]. For initializing memberships, a Bernoulli distribution of parameter $\tau = 0.5$ can be used to generate random memberships. Using either a representative initialization or memberships initialization, the result of the presented methods may be a local optimum of the objective criterion, rather than the global optimum. To deal with this problem, the user should adopt a multi-start procedure by testing different initializations and keeping only the clustering which has the lowest value of the objective criterion.

4 Evaluation of overlapping clustering

The evaluation of clustering, also referred to as cluster validity, is a crucial process to assess the performance of the learning method in identifying relevant groups. This process allows the comparison of several clustering methods and allows the analysis of whether one method is superior to another one. Most of the validity measures traditionally used for clustering assessment, including both internal and external evaluations, become obsolete for overlapping clustering because of the multiple assignment of each observation. Despite this, some works propose an extension of well known validation measures to validate overlapping partitioning. In particular, internal evaluation measures, such as purity and entropy-based measures, cannot capture this aspect of the quality of a given clustering solution because they focus on the internal quality of the clusters. However, external validation measures, essentially Precision-Recall measures, were designed for overlapping partitioning. We give in the following three evaluation methods used for computing precision-recall measures.

4.1 Label based evaluation

Label based evaluation [Tsoumakas et al., 2010] is usually used in the field of Information Retrieval (IR) where each document can discuss several topics. This evaluation method is based on the evaluation of each class separately. Given a set of observations $X = \{x_1, ..., x_N\}$ and two partitions over X to compare, $C = \{c_1, ..., c_k\}$ a non-exclusive partitioning of X into k classes representing true labels, and $\Pi = \{\pi_1, ..., \pi_k\}$ a non-exclusive partitioning of X into k clusters where Π is defined by the clustering algorithm. Known the following:

- True positive TP_{ij}: the number of observations in π_j that exist in c_i,
- False negative FN_{ij}: the number of observations in c_i that not exist in π_j
- False positive FP_{ij}: the number of observations in π_j that not exist in c_i,

the Precision-Recall validation measures are computed for each class i and cluster j as follows:

$$\begin{aligned} Precision_{ij} &= \frac{TP_{ij}}{TP_{ij}+FP_{ij}} \\ Recall_{ij} &= \frac{TP_{ij}}{TP_{ij}+FN_{ij}} \\ F-measure_{ij} &= \frac{(2*Recall_{ij}*Precision_{ij})}{(Recall_{ij}+Precision_{ij})}. \end{aligned}$$

The computation of Precision-Recall measures for all labels is archived using macro-averaging technique which is usually used in Information Retrieval tasks to evaluate clustering results when the number of classes is not large [Yang, 1999] as follows:

$$\begin{aligned} Recall &= \frac{\sum_{i=1}^{k} \max_j Recall_{ij}}{k} \\ Precision &= \frac{\sum_{i=1}^{k} \max_j Precision_{ij}}{k} \\ Fmeasure &= \frac{\sum_{i=1}^{k} \max_j Fmeasure_{ij}}{k}. \end{aligned} \tag{40}$$

4.2 Pair based evaluation

The pair based Precision-Recall measures are calculated over pairs of observations [Banerjee et al., 2005]. For each pair of observations that share at least one cluster in the overlapping clustering results, Precision-Recall measures evaluate whether the prediction of this pair as being in the same cluster is correct with respect to the underlying true class in the data.

Given a set of observation $X = \{x_1, ..., x_N\}$ and two non-exclusive partitionings over X to compare, $C = \{c_1, ..., c_k\}$ a partition of X into k classes, and $\Pi = \{\pi_1, ..., \pi_{k1}\}$ a partition of X into $k1$ clusters and by Considering the following:

- TP : the number of pairs of observations in X that share at least one class in C and share at least one cluster in Π,
- FN : the number of pairs of observations in X that share at least one class in C and do not share any cluster in Π and
- FP : the number of pairs of observations in X that do not share any class in C and share at least one cluster in Π,

the Precision-Recall measures are computed as follows:

$$\text{Precision} = (TP)/(TP+FP)$$
$$\text{Recall} = (TP)/(TP+FN)$$
$$\text{F-measure} = (2*\text{Recall}*\text{Precision})/(\text{Recall}+\text{Precision}).$$

4.3 BCubed evaluation

Given the importance of observation occurrences in clusters and classes in overlapping partitioning, the BCubed evaluation [Amigo et al., 2009] takes into account the multiplicity of classes and clusters which considers the fact that two observations sharing n classes should share n clusters.

BCubed Precision-Recall measures are computed independently for each observation in the partitioning. Let the following:

$$Multiplicity \quad precision(x_i, x_j) = \frac{Min(|\Im(x_i) \cap \Im(x_j)|, |L(x_i) \cap L(x_j)|)}{|\Im(x_i) \cap \Im(x_j)|}$$

$$Multiplicity \quad recall(x_i, x_j) = \frac{Min(|\Im(x_i) \cap \Im(x_j)|, |L(x_i) \cap L(x_j)|)}{|L(x_i) \cap L(x_j)|} \tag{41}$$

where x_i and x_j are two observations, $L(x_i)$ the set of classes and $\Im(x_i)$ the set of clusters associated to observation x_i. In fact, Multiplicity Precision is defined only when the pair of observations (x_i, x_j) share at least one cluster, and Multiplicity Recall is defined only when (x_i, x_j) share at least one class. Multiplicity Precision is maximal, equal to 1, when the number of shared clusters is lower or equal than the number of shared classes and it is minimal, equal to 0, when the two observations do not share any class. Reversely, Multiplicity Recall is maximal when the number of shared classes is lower or equal than the number of shared clusters, and it is minimal when the two observations do not share any cluster.

The BCubed precision associated to one observation will be its averaged multiplicity precision over other observations sharing some of its classes; and the overall BCubed precision will be the averaged precision of all observations. The overall BCubed recall is obtained using the same procedure. The overall BCubed Precision-Recall measures can be formally described by:

$$Precision = AVG_i[AVG_{j.C(x_i)\cap C(x_j)\neq\varnothing} Multiplicity \quad precision(x_i,x_j)]\forall i,j \in \{1,..,N\}$$

$$Recall = AVG_i[AVG_{j.L(x_i)\cap L(x_j)\neq\varnothing} Multiplicity \quad recall(x_i,x_j)]\forall i,j \in \{1,..,N\}$$

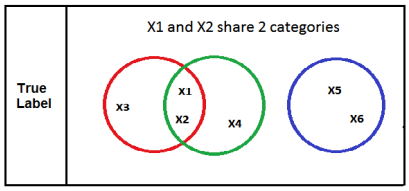
$$F-measure = (2*Precision*Recall)/(Precision+Recall)$$

4.4 Synthesis of evaluation methods for overlapping clustering

Three evaluation methods for assessing the quality of overlapping clustering were presented. The first evaluation method, label based evaluation, is based on matching between labels while the two others are based on pairs of observations. In fact, the label based evaluation requires to label the obtained clusters by matching between classes and clusters which is not a trivial task for unsupervised learning, especially for datasets with large overlaps. The labeling of clusters could lead to biased matching, and consequently lead to biased validation measures. Moreover, the matching between classes and clusters requires to configure a number of clusters equal to that of classes which limits the process of evaluation. Although these limitations, label based evaluation is usually used in Information Retrieval tasks to evaluate clustering results when the number of classes and the sizes of overlaps are not large [Yang, 1999].

To address the issue of labeling the obtained clusters and to make possible the comparison of partitionings with different number of clusters, the pair based Precision-Recall measures are calculated over pairs of observations. This evaluation method offers a flexible evaluation of obtained clusters independently from the real number of classes in the labeled dataset. However, the pair based evaluation has the issue that obtained Recall could be biased as the built overlap in the partitioning decreases and the actual overlap in the dataset increases. The biased Recall is induced by considering only a *binary* function for assessing the relation between a pair of observations and ignoring the *multiplicity* of shared clusters between pairs of observations. For instance, if two observations share three classes in the actual dataset and just share two clusters in the partitioning, the obtained Recall is 1 which is not correct. This problem also occurs for the Precision when actual overlap in the dataset is large.

Given the importance of observation occurrences in clusters and classes in overlapping partitioning, the relation between two observations can not be represented as a binary function. If two observations share two classes and share just one cluster, then the clustering is not capturing completely the relation between both observations as presented in case 3 in Figure 4. On the other hand, if two observations share three clusters but just two classes, then the clustering is introducing more information than necessary as shown in case 3 in Figure 4. This relation is considered for BCubed validation measures by extending the pair based evaluation to take into account the multiplicity of clusters and classes.



(a)

Different cases		Bcubed Evaluation	Pair Evaluation
Case 1	X1 and X2 don't share any cluster	Multiplicity Recall(X1,X2)=min(0,2)/2=0 Multiplicity Precision(X1,X2)=Undefined	Recall(X1,X2)= 0 Precision(X1,X2)= Undefined
Case 2	X1 and X2 share two clusters	MultiplicityRecall(X1,X2)=min(2,2)/2=1 Multiplicity Precision(X1,X2)=min(2,2)/2=1	Recall(X1,X2)= 1 Precision(X1,X2)= 1
Case 3	X1 and X2 share one cluster	MultiplicityRecall(X1,X2)=min(1,2)/2=0.5 MultiplicityPrecision(X1,X2)= min(1,2)/1=1	Recall(X1,X2)= 1 Precision(X1,X2)= 1
Case4	X1 and X2 share three clusters	MultiplicityRecall(X1,X2) =min(3,2)/2=1 MultiplicityPrecision(X1,X2) =min(3,2)/3=0.66	Recall(X1,X2) = 1 Precision(X1,X2) = 1

(b)

Fig. 4 Comparison of Recall and Precision measures computed using the BCubed and the Pair based evaluations: (a) true labels and (b) different cases of the clustering.

Figure 4 shows an illustrative example comparing Precision and Recall computed for a pair of observations (x_1, x_2) using BCubed and pair based evaluations by con-

sidering different case-studies. We notice that Precision and Recall using the label based evaluation are not reported because they can not be computed for a pair of observations. This comparison shows that multiplicity Recall is reduced compared to Recall computed with the pair based evaluation if the partitioning gives less shared clusters than needed as described in case 3. In contrast, if the partitioning gives more shared clusters than actual labels as described in case 4, the multiplicity Precision is reduced compared to the one obtained with pair based evaluation.

As a summary, we can conclude that assessing the quality of overlapping clustering using the BCubed evaluation is more suitable than the pair based evaluation while it takes into account the multiplicity of shared clusters and classes between the pair of observations.

5 Empirical evaluation of overlapping partitional clustering methods

Experiments are performed on real multi-labeled benchmarks from different domains: video classification based on the users ratings (Eachmovie[3] data set), detection of emotion in music songs (Music emotion[4] data set), clustering natural scene image (Scene[5] data set) and clustering on genes (Yeast[6] data set). Table 2 shows the statistics for each data set where "*Labels*" is the number of categories and "*Cardinality*" (natural overlaps) is the average number of categories for each observation. These benchmarks were chosen because of their diversity of application domains, their diversity of sizes ($75 \rightarrow 2417$), their diversity of dimensions ($3 \rightarrow 192$) and their diversity of overlap rates ($1.07 \rightarrow 4.23$).

Results are compared using four validation measures: Precision, Recall, F-measure and Overlap's size. The first three measures are computed using both pair-based and Bcubed-based evaluations as described in Sections 4.2 and 4.3. The fourth measure, Overlap's size, evaluates the size of overlap built by the learning method. This measure is determined by the average number of clusters to which each observation belongs to:

$$Overlap \quad size = \frac{\sum_{i=1}^{N} c_i}{N}, \tag{42}$$

where c_i is the number of clusters to which observation x_i belongs to. The latter is compared with the true rate of overlaps in the labeled data set.

[3] cf. http://www.grouplens.org/node/76.

[4] cf.http://mlkd.csd.auth.gr/multilabel.html

[5] cf.http://mlkd.csd.auth.gr/multilabel.html

[6] cf.http://mlkd.csd.auth.gr/multilabel.html

Table 2 Statistics of used data sets

Data set	Domain	N	Dimension	Labels	Cardinality
EachMovie	Video	75	3	3	1.14
Music emotion	Music	593	72	6	1.86
Scene	Images	2407	192	6	1.07
Yeast	Genes	2417	103	14	4.23

Table 3 Comparison of the performance of overlapping partitioning method in Eachmovie data set

	Pair based evaluation			BCubed Evaluation			Size of
Method	*P.*	*R.*	*F.*	*P.*	*R.*	*F.*	**Overlap**
FCM(θ=0.333, β=2)	0.639	0.675	0.657	0.699	0.683	0.691	**1.12**
ECM (α=1, β=2)	0.566	0.722	0.635	0.575	0.739	0.647	1.32
ECM (α=0.5, β=2)	0.540	0.720	0.617	0.551	0.736	0.631	1.33
OKM	0.465	0.921	0.618	0.399	0.912	0.555	1.70
ALS	0.466	0.855	0.603	0.366	0.819	0.506	1.73
KOKMϕ (RBF)	0.660	0.757	**0.705**	0.685	0.735	**0.709**	1.20
KOKMϕ (Polynomial)	0.459	0.391	0.552	0.416	0.675	0.514	1.48
P. R-OKM ($\alpha = 1$)	0.691	0.731	**0.711**	0.741	0.715	**0.728**	**1.13**
P. R-OKM ($\alpha = 0.1$)	0.505	0.909	0.65	0.440	0.889	0.588	1.57

Table 4 Comparison of the performance of overlapping partitioning method in Emotion data set

	Pair based evaluation			BCubed Evaluation			Size of
Method	*P.*	*R.*	*F.*	*P.*	*R.*	*F.*	**Overlap**
FCM(θ=0.166, β=2)	0.492	0.394	0.437	0.480	0.354	0.408	1.43
ECM (α=1, β=2)	0.482	0.675	0.562	0.373	0.648	**0.473**	2.61
ECM (α=0.5, β=2)	0.488	0.703	0.576	0.360	0.711	**0.478**	2.80
OKM	0.483	0.646	0.552	0.353	0.544	0.428	2.35
ALS	0.471	0.999	**0.640**	0.307	0.970	**0.466**	3.46
KOKMϕ (RBF)	0.487	0.548	0.516	0.396	0.462	0.426	2.15
KOKMϕ (Polynomial)	0.481	0.360	0.412	0.446	0.288	0.352	1.52
P. R-OKM ($\alpha = 1$)	0.493	0.351	0.410	0.86	0.278	0.354	1.26
P. R-OKM ($\alpha = 0.1$)	0.487	0.569	0.525	0.392	0.475	**0.430**	**1.88**

Tables 3, 4, 5 and 6 report Precision (P.), Recall (R.), F-measure (F.) and Size of overlaps for the best run of each method, which gives the minimal value of the objective criterion, among twenty runs on Eachmovie, Emotion, Scene and Yeast data sets. Results of some methods in Yeast and Scene data sets are not reported because of their computational complexity which becomes time consuming and need more than 24 hours. A different initialization of prototypes have been used over the twenty runs, whereas within each run the same initialization of prototypes have been used for the different methods. We note that the number of clusters considered within each method was set to the underlying number of labels in each data set. For the other parameters we consider the following:

Table 5 Comparison of the performance of overlapping partitioning method in Scene data set

	Pair based evaluation			BCubed Evaluation			Size of
Method	*P.*	*R.*	*F.*	*P.*	*R.*	*F.*	**Overlap**
FCM(θ=0.166, β=2)	0.247	0.946	0.392	0.102	0.946	0.185	3.34
ECM (α=1, β=2)	0.255	0.848	0.393	0.110	0.908	0.171	2.96
ECM (α=0.5, β=2)	0.285	0.813	**0.422**	0.105	0.879	0.188	3.01
OKM	0.233	0.928	0.372	0.192	0.926	0.216	2.85
ALS	-	-	-	-	-	-	-
KOKMϕ (RBF)	0.247	0.813	0.379	0.191	0.809	0.309	2.10
KOKMϕ (Polynomial)	0.236	0.844	0.369	0.156	0.848	0.263	2.36
P. R-OKM ($\alpha = 1$)	0.467	0.426	**0.446**	0.490	0.438	**0.462**	**1.00**
P. R-OKM ($\alpha = 0.1$)	0.30	0.791	**0.435**	0.293	0.797	**0.428**	**1.69**

Table 6 Comparison of the performance of overlapping partitioning method in Yeast data set

	Pair based evaluation			BCubed Evaluation			Size of
Method	*P.*	*R.*	*F.*	*P.*	*R.*	*F.*	**Overlap**
FCM(θ=0.0714, β=2)	0.784	1.00	**0.879**	0.148	1.00	0.257	13.60
ECM (α=1, β=2)	-	-	-	-	-	-	-
ECM (α=0.5, β=2)	-	-	-	-	-	-	-
OKM	0.783	0.877	**0.827**	0.587	0.485	**0.531**	**4.80**
ALS	-	-	-	-	-	-	-
KOKMϕ (RBF)	0.785	0.793	**0.789**	0.558	0.467	**0.509**	5.03
KOKMϕ (Polynomial)	0.782	0.755	0.768	0.614	0.398	0.483	**4.59**
P. R-OKM ($\alpha = 1$)	0.801	0.075	0.137	0.806	0.014	0.027	1.00
P. R-OKM ($\alpha = 0.1$)	0.783	0.546	0.643	0.749	0.178	0.287	3.04

- for FCM, we set the fuzziness parameter $\beta = 2$ and the threshold value $\theta = 1/k$ with k denotes the number of clusters. However, we note that obtained results of FCM show a high sensitivity to the used threshold as well as the number of clusters increases: for example, in the Yeast data set, using a threshold equal to 0.0714 the obtained F-measure is equal to 0.257 while using the same method with a threshold equal to 0.0715 the F-measure decreases to 0.017 because clusters memberships are almost null. These results show the limit of the extension of FCM to detect overlapping groups.
- for ECM, we fix the fuzziness parameter $\beta = 2$ and we test two values of α equal to 1 and 0.5 for all the data sets.
- for KOKMϕ, we perform different executions using two types of kernels which are RBF (Radial basis Function) and polynomial kernels with parameters $\sigma = 1000$ and $d = 3$ respectively. As for FCM, we show that results are highly sensitive to the type and the considered parameters of the kernel.
- for Parameterized R-OKM, we perform experiments using two values of α equal to 1 and 0.1 for all the data sets. We note that other considered values of α largely improves the performance of Parameterized R-OKM. However, we only report these two values of α to standardize the presented results for the different benchmarks.

- for ALS, reported results are built after reducing data in the interval [-1 1]. We note that ALS fails to build overlapping groups in almost benchmarks (Eachmovie, Emotion and Scene) when data are not centered to have zero mean.

The analysis of the experimental results firstly shows the reliability of the extended Bcubed measures for evaluating overlapping clustering compared to pair based-evaluation: obtained F-measures are higher for clusterings whose overlap rate comes closer to the expected one on the whole. For example, in Scene data set (actual overlap $= 1.07$) using pair based-evaluation obtained F-measures for ECM and Parameterized R-OKM are nearly equal (0.442 and 0.435 respectively). The F-measure obtained with ECM is induced by high Recall since overlaps rate (overlap$= 3$) largely exceeds the actual overlap in Scene while for Parameterized R-OKM F-measure is induced by high Precision since built overlap (overlap=1.68) is near to the actual one. However, using Bcubed evaluation we obtain more reliable evaluation where clustering obtained with Parameterized R-OKM largely exceeds clustering obtained with ECM in terms of F-measure.

Second, empirical results show that Overlap rates have a strong influence on the matching measurement of the clusterings with respect to the expected classes. Methods which allow to control overlap's size, such as Parameterized R-OKM, ECM and FCM can give partitionings which better fit existing structures in the data set, unless good customization of their parameters is required. Whatever the approach used, almost of evaluated methods produce large overlaps exceeding the expected rates, known the actual ones. Producing partitionings which have regulated overlaps improves the F-measure and leads to non-disjoint groups fitting better the data. This fact demonstrates that size of overlaps is an important characteristic that should be controlled while building overlapping groups.

The third conclusion concerns the comparison of uncertain and hard memberships based methods. The results show the limit of using uncertain memberships methods when the number of expected clusters increases. As the case for Yeast data set, uncertain memberships methods require a 10^{-4} precision to fix a threshold value for FCM to be able to produce a non disjoint partitioning of data. However, overlapping groups are easily built using hard memberships methods. We also notice the importance of the computational complexity of overlapping clustering methods when the size of data become large either in terms of number of observations or expected number of clusters.

6 Conclusion

We focused in this paper on overlapping clustering, for which we give a classification of existing methods based on the conceptual approach to look for non-disjoint partitioning. Our study is essentially based on overlapping methods which are based on the partitional approach. For that, we survey existing overlapping partitional methods in the literature, classified in two main categories: uncertain memberships and hard memberships. We also gave theoretical and experimental comparisons of

existing partitional methods. Furthermore, another important issue that we discussed in this paper is overlapping cluster validity. We presented three external methodologies used to evaluate reliability of non-disjoint partitionings through the evaluation of Precision-recall measures.

At the end of this survey, we claim that overlapping clustering has a growing interest in machine learning research while many real life applications require a non disjoint partitioning of data. Many active challenges in overlapping clustering applications motivate researchers to propose more perfective and efficient learning process. For example, recent works are interested with the identification of non disjoint groups when data contain outliers, or detecting overlapping clusters when data have groups with uneven density. Other works are interested with the identification of overlapping clustering from data having high number of dimensions, or a huge number of observations. All these challenges within overlapping clustering open an exciting directions for future researchers.

References

[Amigo et al., 2009] Amigo, E., Gonzalo, J., Artiles, J., and Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486.

[Banerjee et al., 2005] Banerjee, A., Krumpelman, C., Basu, S., Mooney, R. J., and Ghosh, J. (2005). Model based overlapping clustering. In *International Conference on Knowledge Discovery and Data Mining*, pages 532–537.

[Baumes et al., 2005] Baumes, J., Goldberg, M., and Magdon-Ismail, M. (2005). Efficient identification of overlapping communities. In *IEEE international conference on Intelligence and Security Informatics*, pages 27–36.

[BenN'Cir et al., 2013] BenN'Cir, C., Cleuziou, G., and Essoussi, N. (2013). Identification of non-disjoint clusters with small and parameterizable overlaps. In *IEEE International Conference on Computer Applications Technology (ICCAT)*, pages 1–6.

[BenN'Cir and Essoussi, 2012] BenN'Cir, C. and Essoussi, N. (2012). Overlapping patterns recognition with linear and non-linear separations using positive definite kernels. *International Journal of Computer Applications (IJCA)*, pages 1–8.

[BenN'Cir et al., 2010] BenN'Cir, C., Essoussi, N., and Bertrand, P. (2010). Kernel overlapping k-means for clustering in feature space. In *International Conference on Knowledge discovery and Information Retrieval (KDIR)*, pages 250–256.

[Berkhin, 2006] Berkhin, P. (2002). A survey of clustering data mining techniques. In *Grouping Multidimensional Data*, pages 25–71.

[Bertrand and Janowitz, 2003] Bertrand, P. and Janowitz, M. (2003). The k-weak hierarchical representations: an extension of the indexed closed weak hierarchies. *Discrete Applied Mathematics*, 127(2):199–220.

[Bezdek, 1981] Bezdek, J. C. (1981). Pattern recognition with fuzzy objective function algoritms. *Plenum Press*, 4(2):67–76.

[Bonchi et al., 2011] Bonchi, F., Gionis, A., and Ukkonen, A. (2011). Overlapping correlation clustering. In *11th IEEE International Conference on Data Mining (ICDM)*, pages 51–60.

[Bonchi et al., 2013] Bonchi, F., Gionis, A., and Ukkonen, A. (2013). Overlapping correlation clustering. *Knowledge and Information Systems*, 35(1):1–32.

[Celebi and Kingravi, 2012] Celebi, M.-E, Kingravi, H.(2012). Deterministic initialization of the k-means algorithm using hierarchical.clustering *International Journal of Pattern Recognition and Artificial Intelligence*, 26(7):1250018.

[Celebi et al., 2013] Celebi, M.-E, Kingravi, H., and Vela, P.-A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, 40(1):200–210.

[Cleuziou, 2008] Cleuziou, G. (2008). An extended version of the k-means method for overlapping clustering. In *International Conference on Pattern Recognition ICPR*, pages 1–4.

[Cleuziou, 2009] Cleuziou, G. (2009). Two variants of the okm for overlapping clustering. *Advances in Knowledge Discovery and Management*, pages 149–166.

[Cleuziou, 2013] Cleuziou, G. (2013). Osom: A method for building overlapping topological maps. *Pattern Recognition Letters*, 34(3):239–246.

[Davis and Carley, 2008] Davis, G. B. and Carley, K. M. (2008). Clearing the fog: Fuzzy, overlapping groups for social networks. *Social Networks*, 30(3):201–212.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

[Depril et al., 2012] Depril, D., Mechelen, I. V., and Wilderjans, T. F. (2012). Lowdimensional additive overlapping clustering. *Journal of Classification*, 29(3):297–320.

[Depril et al., 2008] Depril, D., Van Mechelen, I., and Mirkin, B. (2008). Algorithms for additive clustering of rectangular data tables. *Computational Statistics and Data Analysis*, 52(11):4923–4938.

[Diday, 1984] Diday, E. (1984). Orders and overlapping clusters by pyramids. Technical Report 730, INRIA, France.

[Fellows et al., 2011] Fellows, M. R., Guo, J., Komusiewicz, C., Niedermeier, R., and Uhlmann, J. (2011). Graph-based data clustering with overlaps. *Discrete Optimization*, 8(1):2–17.

[Fu and Banerjee, 2008] Fu, Q. and Banerjee, A. (2008). Multiplicative mixture models for overlapping clustering. In *8th IEEE International Conference on Data Mining*, pages 791–796.

[Gil-García and Pons-Porrata, 2010] Gil-García, R. and Pons-Porrata, A. (2010). Dynamic hierarchical algorithms for document clustering. *Pattern Recogn. Lett.*, 31(6):469–477.

[Goldberg et al., 2010] Goldberg, M., Kelley, S., Magdon-Ismail, M., Mertsalov, K., and Wallace, A. (2010). Finding overlapping communities in social networks. In *IEEE Second International Conference on Social Computing (SocialCom)*, pages 104–113.

[Gregory, 2007] Gregory, S. (2007). An algorithm to find overlapping community structure in networks. In *Knowledge Discovery in Databases: PKDD 2007*, volume 4702, pages 91–102.

[Gregory, 2008] Gregory, S. (2008). A fast algorithm to find overlapping communities in networks. In *Machine Learning and Knowledge Discovery in Databases*, volume 5211, pages 408–423.

[Halkidi et al., 2001] Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2-3):107–145.

[Heller and Ghahramani, 2007] Heller, K. and Ghahramani, Z. (2007). A nonparametric bayesian approach to modeling overlapping clusters. In *AISTATS*.

[Jain, 2010] Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666.

[Krishnapuram and Keller, 1993] Krishnapuram, R. and Keller, J. M. (1993). A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, 1(2):98–110.

[Lingras and West, 2004] Lingras, P. and West, C. (2004). Interval set clustering of web users with rough k-means. *Journal of Intelligent Information System*, 23(1):5–16.

[Liu et al., 2012] Liu, Z.-G., Dezert, J., Mercier, G., and Pan, Q. (2012). Belief c-means: An extension of fuzzy c-means algorithm in belief functions framework. *Pattern Recognition Letters*, 33(3):291–300.

[MacQueen, 1967] MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297.

[Magdon-Ismail and Purnell, 2011] Magdon-Ismail, M. and Purnell, J. (2011). Ssde-cluster: Fast overlapping clustering of networks using sampled spectral distance embedding and gmms. In *IEEE third international conference on social computing (socialcom)*, pages 756–759.

[Masson and Denoeux, 2008] Masson, M.-H. and Denoeux, T. (2008). Ecm: An evidential version of the fuzzy c-means algorithm. *Pattern Recognition*, 41(4):1384 – 1397.

[Mirkin, 1987] Mirkin, B. G. (1987). Method of principal cluster analysis. *Automation and Remote Control*, 48:1379–1386.

[Mirkin, 1990] Mirkin, B. G. (1990). A sequential fitting procedure for linear data analysis models. *Journal of Classification*, 7(2):167–195.

[Pérez-Suárez et al., 2013a] Pérez-Suárez, A., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., and Medina-Pagola, J. E. (2013a). Oclustr: A new graph-based algorithm for overlapping clustering. *Neurocomputing*, 109(0):1–14.

[Pérez-Suárez et al., 2013b] Pérez-Suárez, A., Martnez-Trinidad, J. F., Carrasco-Ochoa, J. A., and Medina-Pagola, J. E. (2013b). An algorithm based on density and compactness for dynamic overlapping clustering. *Pattern Recognition*, 46(11):3040–3055.

[Richard Duda, 2001] Richard Duda, Peter Hart, D. S. (2001). *Pattern Classification*.

[Snoek et al., 2006] Snoek, C. G. M., Worring, M., van Gemert, J. C., Geusebroek, J.-M., and Smeulders, A. W. M. (2006). The challenge problem for automated detection of 101 semantic concepts in multimedia. In *14th annual ACM international conference on Multimedia*, pages 421–430.

[Tang and Liu, 2009] Tang, L. and Liu, H. (2009). Scalable learning of collective behavior based on sparse social dimensions. In *ACM conference on Information and knowledge management*, pages 1107–1116.

[Tsoumakas et al., 2010] Tsoumakas, G., Katakis, I., and Vlahavas, I. (2010). Mining Multi-label Data. In *Data Mining and Knowledge Discovery Handbook*, pages 667–685.

[Wang and Fleury, 2011] Wang, Q. and Fleury, E. (2011). Uncovering overlapping community structure. In *Complex Networks*, volume 116, pages 176–186.

[Wang et al., 2010] Wang, X., Tang, L., Gao, H., and Liu, H. (2010). Discovering overlapping groups in social media. In *IEEE International Conference on Data Mining*, pages 569–578.

[Wieczorkowska et al., 2006] Wieczorkowska, A., Synak, P., and Ras, Z. (2006). Multi-label classification of emotions in music. In *Intelligent Information Processing and Web Mining*, volume 35 of *Advances in Soft Computing*, pages 307–315.

[Wilderjans et al., 2011] Wilderjans, T., Ceulemans, E., Mechelen, I., and Depril, D. (2011). Adproclus: A graphical user interface for fitting additive profile clustering models to object by variable data matrices. *Behavior Research Methods*, 43(1):56–65.

[Wilderjans et al., 2013] Wilderjans, T. F., Depril, D., and Mechelen, I. V. (2013). Additive biclustering: A comparison of one new and two existing als algorithms. *Journal of Classification*, 30(1):56–74.

[Yang, 1999] Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*, 1:67–88.

[Zhang et al., 2007a] Zhang, S., Wang, R.-S., and Zhang, X.-S. (2007a). Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A: Statistical Mechanics and its Applications*, 374(1):483–490.