Generalization of c-means for identifying non-disjoint clusters with overlap regulation

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

Generalization of *c*-means for identifying non-disjoint clusters with overlap regulation ☆

Chiheb-Eddine ben N'Cir [a,*], Guillaume Cleuziou [b,c,*], Nadia Essoussi [a]

[a] *LARODEC, ISG Tunis, Université de Tunis, 41 Avenue de la liberté, cité Bouchoucha, 2000 Le Bardo, Tunisia*
[b] *Univ. Orléans, INSA Centre Val de Loire, LIFO EA 4022, FR-45067 Orléans, France*
[c] *GREYC, Université de Caen Basse-Normandie, Campus Côte de Nacre, Boulevard du Maréchal Juin, BP 5186, 14032 CAEN Cedex, France*

Keywords:
Overlapping clustering
Overlap regulation
Clustering evaluation
Multi-label data

A B S T R A C T

Clustering is an unsupervised learning method that enables to fit structures in unlabeled data sets. Detecting overlapping structures is a specific challenge involving its own theoretical issues but offering relevant solutions for many application domains. This paper presents generalizations of the *c*-means algorithm allowing the parametrization of the overlap sizes. Two regulation principles are introduced, that aim to control the overlap shapes and sizes as regard to the number and the dispersal of the cluster concerned. The experiments performed on real world datasets show the efficiency of the proposed principles and especially the ability of the second one to build reliable overlaps with an easy tuning and whatever the requirement on the number of clusters.

1. Introduction

Clustering is an important task in data mining that aims to organize a dataset into groups (or clusters) containing similar data. Clustering is used successfully in many fields of application such as: marketing, to find groups of customers with similar purchasing behaviors; biology, to group plants or animals into clusters of similar species or Information Retrieval to organize documents with similar topics into groups. In many real world applications, data naturally organize themselves into non-disjoint groups thus requiring to find a coverage of the data with overlapping clusters rather than a partition. The corresponding research domain has been referred to as *overlapping clustering* and studied through various approaches during the last decades [28,11,2,10,12].

Overlapping clustering methods contribute to solve many real life problems which require the assignment of each observation to multiple clusters: in social network analysis, community extraction algorithms should be able to detect overlapping clusters because an actor can belong to multiple communities [31,33,12]; in video classification, each entry can potentially have multiple genres [29]; in emotion detection, clustering algorithms should be able to raise several emotions for a specific piece of music [34] and Information Retrieval systems should also be able to group documents with several topics into several corresponding clusters [14,26], etc.

Unlike *fuzzy clustering*, overlapping clustering assumes that an observation can really belong to several clusters (without any membership degree). Whatever the approach used (hierarchical or criterion-based), existing algorithms produce clusterings without possibility of control on the size (or quality) of the overlaps, other than the constraints on the overlaps set by the model itself. Although the method ideally should reveal the clustering that best fit to the data, such a *best* clustering is usually not unique and the overlapping dimension must be used as a parameter of the pattern recognition process.

The present paper is based on the well known *c*-means algorithm and introduces two new methods R_1-OKM and R_2-OKM, both generalizing the initial model OKM (*Overlapping k-means*) and allowing a regulation of the overlaps. The two regulation principles aim to measure the benefits of an overlap with respect to the number and the dispersal of the concerned clusters respectively.

The next section is devoted to a description of the works related to overlapping clustering and the background required about the OKM model. Then, the overlap regulation principle and the two models R_1-OKM and R_2-OKM are detailed in Section 3. In Section 4 we discuss and propose an evaluation process for overlapping clustering algorithms and we assess the positioning of the new models with respect to the other competitive methods. Finally, Section 5 reports on the conclusions and the perspectives this study opens up.

☆ This paper has been recommended for acceptance by S. Todorovic.

* Corresponding authors. Tel.: +216 98626731 (C.-E. ben N'Cir). Address: Univ. Orléans, INSA Centre Val de Loire, LIFO EA 4022, FR-45067 Orléans, France. Tel.: +33 238492591; fax: +33 2 38 41 71 37 (G. Cleuziou).

E-mail addresses: chiheb.benncir@isg.rnu.tn (C.-E. ben N'Cir), guillaume.cleuziou@univ-orleans.fr (G. Cleuziou), nadia.essoussi@isg.rnu.tn (N. Essoussi).

2. Preliminaries

2.1. Related works

The overlapping clustering problem has been studied since the last four decades. Two kinds of approaches have been led: the *heuristical* and *theoretical* solutions.

We denoted as *heuristical* the solutions that consist either in modifying the clusters resulting from a standard clustering algorithm into overlapping clusters (typically results from *c*-means or fuzzy-*c*-means algorithms [19,35]) or in proposing new clustering solutions based on intuitive processes which lead to build overlaps but they are not based on any overlap model; the CBC algorithm (*Clustering by Committee*) introduced by Pantel and Lin [24] for textual data clustering and the POBOC algorithm (*Pole-Based Overlapping Clustering*) proposed by Cleuziou et al. [7] are two distinctive examples of such intuitive processes that generate overlapping clusters. These contributions lead to suitable results in some contexts but they are not predicated on theoretical models and their extension or improvement are limited as a rule.

Conversely, the second kind of approaches (theoretical solutions) are extensions from traditional clustering models such as hierarchical, generative, graph-based or reallocation methods. The overlapping variants of hierarchies are the pyramids [11] and more generally the weak-hierarchies [4]. Concisely, these variants aim to reduce the discrepancy between the original dissimilarities over the considered dataset and the ones induced by the (pseudo)-hierarchical structure. However, these structures are either restrictive on the overlaps (as for pyramids) or hard to build and visualize (as for weak-hierarchies).

Overlapping methods based on graph theory are mostly used in the context of community detection in complex networks [3,16,35,17,8,15,32,12,21]. This category of methods models the initial collection of observations as a thresholded similarity graph which can be directed or undirected graph depending on the problem. These methods build non-disjoint partitioning of vertices (observations) through the cover of the main graph by its decomposition to subgraphs using special techniques like the star shaped sub-graph (s-graph). The difference between them consists on the criterion used for ordering and selecting the sub-graphs. The obtained overlaps are usually natural intersections of vertices belonging to multiple subgraphs. This fact explains the shortcoming of high overlaps between clusters that graph based methods suffer from. Another shortcoming of these methods is the computational complexity which is usually exponential in the number N of vertices and could be reduced to $O(N^2)$ as the case for OClustR (*Overlapping Clustering based on Relevance*) [25].

Overlapping clustering approaches using generative mixture models [2,18,13] have been proposed as extensions of the EM (*Expectation Maximization*) algorithm [9]. These models are mainly supported by biological processes; they hypothesize that each data is the result of a mixture of distributions: the mixture can be additive [2] or multiplicative [18,13] and the probabilistic framework make 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.

We focus our study on another type of overlapping clustering methods, formalized through objective criteria to optimize and solved with usual reallocation processes. Two types of models have been proposed that refer to different hypothesis on overlapping modelling:

- The *additive modelling*, introduced initially by Shepard and Arabie [28] and took over by Mirkin [23] and then by Banerjee et al. [2] (in a model-based formalization) and Depril et al. [10], hypothesizes that overlaps result in the addition of the profiles of the related clusters. Additive models are successfully applied in various application domains like marketing, gene expression and psychology for which a data sharing several behaviors can actually be modeled by an additive combination of their profiles.
- We denote, conversely, as *geometrical*, the models that formalize an overlap as a mean (or barycentre) on the related cluster profiles. They are based on a geometrical reasoning in the data space and they match with many real world multi-labelled data sets, as shown in the following. The first geometrical models have been proposed by Cleuziou [6] and Masson and Denoeux [22].

We detail, in the following of this section, the algorithms ALS [10] and OKM [6] and their underlying additive and geometrical model respectively.

2.2. ALS vs. OKM: models and algorithms

Given a data matrix $X = (x_1, \ldots, x_N)^T$ with N observations in $\mathbb{R}^M$, the Alternating Least Square algorithm (ALS) consists in minimizing the following sum of local errors:

$$J_{ALS}(A,P) = \sum_{i=1}^{N}\sum_{j=1}^{M}\left(x_{i,j} - \sum_{k=1}^{K} a_{i,k}p_{k,j}\right)^2, \tag{1}$$

where K is a given number of expected clusters, A is a binary $(N \times K)$ assignment matrix ($a_{i,k} = 1$ if x_i belongs to the k^{th} cluster) and P is the $\mathbb{R}^{K\times M}$ matrix of cluster profiles. As an additive model, ALS defines a local error for each data x_i by its (square) euclidean distance with the sum of its cluster profiles.

Similarly, the objective function used in OKM is based on local errors but differs on the combination of cluster profiles. As a geometrical model, OKM aims to match each data with the mean (barycentre) of its cluster profiles:

$$J_{OKM}(A,P) = \sum_{i=1}^{N}\sum_{j=1}^{M}\left(x_{i,j} - \frac{\sum_k a_{i,k}p_{k,j}}{\sum_k a_{i,k}}\right)^2. \tag{2}$$

Both models can be formulated as matrix decomposition problems trying to approximate X either by AP for ALS or SP (with $s_{i,k} = \frac{a_{i,k}}{\sum_l a_{i,l}}$) for OKM. The corresponding objective functions can be rewritten using the Frobenius norm $\|.\|_F$ as:

$$J_{ALS} = \|X - AP\|_F^2 \text{ and } J_{OKM} = \|X - SP\|_F^2, \tag{3}$$

and minimized by a common iterative reallocation process as described in Algorithm 1 which alternates (1) update of the assignments (A or S) and (2) update of the cluster profiles (P):

Algorithm 1. Two-step reallocation process for overlapping clustering

Require: X: a data set described over $\mathbb{R}^M$.
$J(.)$: an objective function.
K: number of clusters.
Ensure A: binary membership matrix and P: matrix of cluster profiles
Initialize cluster profiles P, randomly picking K data in X
do
1: Compute new cluster memberships A with P fixed
2: Compute new cluster profiles P with A fixed
while $J(A,P)$ decreases
Return the final cluster membership matrix A.

Algorithm 2. OKM assignment strategy

Require: x_i: a data vector to assign.
$\tilde{A}_i$: the previous assignment for x_i.
$J(.)$: The objective function of OKM.
P: a matrix of K cluster profiles.
Ensure A_i: a binary membership vector
Look for the index q of the nearest cluster profile of x_i:
$q = \arg\min_{k\in[[1;K]]}\|x_i - A_i P_{k,.}\|$
Set $A_{i,k} = 0, \forall k \neq q$ and set $A_{i,q} = 1$
do
Look for the index q of the nearest cluster profile of x_i:
$q = \arg\min_{k\in[[1;K]]\wedge A_{i,k}=0}\|x_i - A_i P_{k,.}\|$
$A'_i = A_i$ and $A'_{i,q} = 1$
if$(J(A'_i, P) \leqslant J(A_i, P))$ **then** $A_i = A'_i$
else $q = 0$
while $(q > 0$ and $\|A_i\| < K)$
If $(J(A_i, P) \leqslant J(\tilde{A}_i, P))$ **then return** A_i
Else return $\tilde{A}_i$

1. The assignment step is a non-trivial discrete optimization problem that can be solved as in ALS by considering for each data x_i any of the 2^K combinations or with approximation heuristics avoiding the combinatorial problem, as proposed in OKM or MOC [2].
2. To update the cluster profiles, an exact solution P^* can be obtained by considering the pseudo-inverse of A (or S): $P^* = (A^T A)^{-1} A^T X$ as proposed in ALS and MOC. OKM performs updates for each cluster successively thus leading to non-optimal profiles (cf. Algorithm 2) but avoiding the (costly) matrix inversion.

The two algorithms ALS and OKM proceed similarly but their underlying model are not compatible. As mentioned above, the choice between an additive or a geometrical model depends on the data structure. One can show that when emerging clusters are properly distributed around the gravity centre of the data space, a preprocessing on the data (at least centering) makes the additive models to proceed almost as geometrical models. It appears from previous studies, that the success of additive models most often requires such a preprocessing thus suggesting that a geometrical model would be more appropriate.

In the following, the problem of overlap regulation for clustering approaches is considered for the OKM geometrical model, without loss of generality since similar regulation principles could be performed on additive models.

3. Overlap regulations

Into a knowledge discovery process, the user or expert is central and must have the means to interact with the system; as well as he examines several possibilities on the number of clusters, the metric to use or the fuzziness of the solution in a clustering process, he must be able to regulate the size of the overlaps when such an overlapping structuring is expected.

To make the overlapping regulation feasible for geometrical model we formalize two regulation principles that are based respectively on the number and the dispersal of the cluster profiles concerned by the data assignment. Instead of considering strong global thresholds that could for example limit or favor for any data the number or dispersal of its memberships to a given value regardless of the data context, the regulation process we propose is data sensitive; it aims to parameterize the expected benefit of an overlap on the local errors.

To give the reader a visual reading of the regulation principles, we illustrate the overlaps with Voronoï cells on a two-dimensional space with three clusters. Fig. 1a shows the Voronoï cells provided by the OKM model: three clusters profiles are considered with coordinates $(4,8)$, $(2,6)$ and $(8,3)$ and colors *blue*, *red* and *yellow* respectively; each color area is a voronoï cell that circumscribes one cluster or one possible intersection (overlap) of clusters. As an example, any data located into the *yellow* area will be assigned only to cluster 3 with the OKM model, while data located into the *green* area will be assigned to the overlap between clusters 1 and 3 ($blue \wedge yellow \rightarrow green$) and the *black* area illustrates the overlap between the three clusters.

3.1. Overlap regulation based on the number of assignments (R_1-OKM)

Additive and geometrical models favor the assignment of a data x_i to the nearest[1] combination of cluster profiles, as described by their objective criteria (1) and (2), regardless of the number of cluster profiles. Let A_i and A'_i two cluster assignment combinations relative to x_i represented by binary vectors into $\{0,1\}^K$, the decision of assigning a data x_i to the combination A_i rather than to A'_i in OKM requires a positive difference in the induced local errors. We propose a first overlap regulation principle R_1-OKM by weighting the local errors with the size of the combinations adjusted by a parameter α. The decision of the assignment is now regulated by the following expression:

$$\|A'_i\|^{\alpha}.\|x_i - p^{A'_i}\|^2 - \|A_i\|^{\alpha}.\|x_i - p^{A_i}\|^2 > 0, \tag{4}$$

where p^{A_i} denotes the combination $p_j^{A_i} = \sum_k a_{i,k}p_{k,j}/\|A_i\|$. As an example, for $\alpha = 1$, $\|A_i\| = 2$ and $\|A'_i\| = 1$, the new model favors the assignment of x_i to the combination A_i only if the local error induced is twice as lower as the local error induced with $A'_i\left(\left(\frac{\|A_i\|}{\|A'_i\|}\right)^{\alpha} = 2\right)$. Such an overlap regulation model results in the following objective criterion:

$$J_{R_1 OKM}(A,P) = \sum_i \left[\sum_k a_{i,k}\right]^{\alpha} \sum_j \left(x_{i,j} - \frac{\sum_k a_{i,k}p_{k,j}}{\sum_k a_{i,k}}\right)^2 \tag{5}$$

Let us notice the behavior of the model, depending on the values of α : $\alpha = 0$ annihilates the weighting on the combination sizes thus leading to the original OKM model; $\alpha > 0$ penalizes the assignments to wide combinations as well as α increases until a non-overlapping model identical to the usual (c-means) least-square model ($\alpha \rightarrow +\infty$); $\alpha < 0$ favors the overlaps as well as α decreases until a trivial clustering scheme with any data assigned to any clusters ($\alpha \rightarrow -\infty$).

Fig. 1b and c illustrate the assignment behavior of the R_1-OKM model with $\alpha = 1$ and $\alpha = -1$ respectively. One can see the non-linear cluster boundaries provided by the new model and the expected behavior of the model, with smaller overlaps when $\alpha > 0$ and greater overlaps when $\alpha < 0$.

The optimization algorithm for R_1-OKM follows the general reallocation process previously described in Algorithm 1. The assignment strategy remains (cf. Algorithm 2) but using the new objective criterion $J_{R_1 OKM}$. The updating of cluster profiles is performed successively for each cluster and requires to derivate the optimal cluster profile $P^*_{k,.}$ given the assignment matrix A and the rest of the matrix P: we obtain that such an optimal cluster profile is defined by:

$$p^*_{k,j} = \frac{\sum_i a_{i,k}\left(\sum_l a_{i,l}\right)^{\alpha-2} p^i_{k,j}}{\sum_i a_{i,k}\left(\sum_l a_{i,l}\right)^{\alpha-2}}, \tag{6}$$

[1] In the sense of euclidean distance.

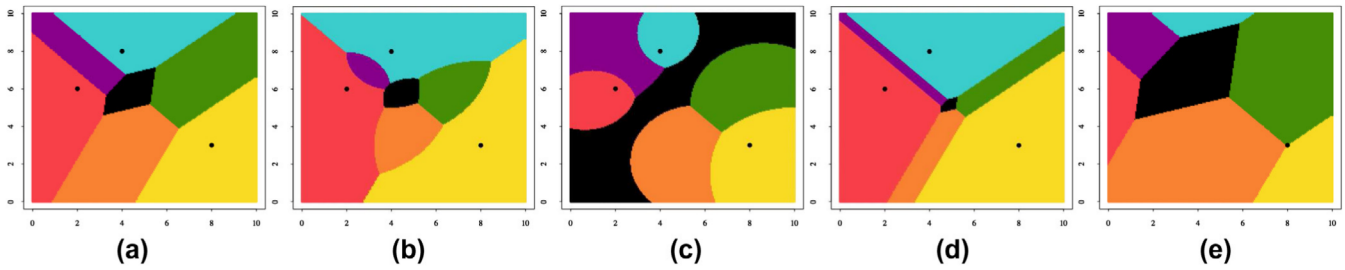

Fig. 1. Two-dimensional example of Voronoï cells induced by the original OKM model (a), the new regulated R_1-OKM model with parameter $\alpha = 1$ (b) and $\alpha = -1$ (c) and the new regulated R_2-OKM model with parameter $\lambda = 2$ (d) and $\lambda = -0.5$ (e).

with notation P_k^i denoting the perfect cluster profile P_k according to x_i, such that the local error on x_i is zero: $p^i_{k,j} = x_{i,j}\sum_l a_{i,l} - \sum_{l\neq k} a_{i,l}p_{l,j}$. Thus, the new cluster profile $P^*_{k.}$ is a weighted average on the cluster profiles expected by each data assigned to cluster k.

3.2. Overlap regulation based on the dispersal of the clusters (R_2-OKM)

We introduce a second model that uses the dispersal of the cluster profiles from the data to assign, in order to regulate the overlap significance. The new R_2-OKM model hypotheses that the overlap regulation must works differently depending on the region in the cluster space; more precisely we assume that (for example) when the expert aims to reduce cluster overlaps, his first expectation is to separate more the distant clusters. To formalize such a regulation principle, we propose to favor or penalize the local error relative to each data x_i by using the following dispersal criterion quantifying the average (square) distance from x_i to its cluster profiles:

$$D_{A,P}(x_i) = \frac{\sum_k a_{i,k}\sum_j (x_{i,j} - p_{k,j})^2}{\sum_k a_{i,k}}. \tag{7}$$

Based on the later principle of overlap regulation, the new objective criterion of R_2-OKM is defined by:

$$J_{R_2OKM}(A,P) = \sum_i \sum_j \left[\left(x_{i,j} - \frac{\sum_k a_{i,k}p_{k,j}}{\sum_k a_{i,k}} \right)^2 + \lambda . D_{A,P}(x_i) \right] \tag{8}$$

where λ denotes the parameter allowing to control the dispersal criterion and plays a role similar to α in the first model; $\lambda = 0$ annihilates the regulation leading to OKM, $\lambda > 0$ restrains overlaps while $\lambda < 0$ favors more overlaps. Fig. 1 illustrates the (linear) shapes of the overlap boundaries induced by the new regulation principle: in Fig. 1d ($\lambda = 2$), the overlap limitation is strongest in proportion with cluster 3 (*yellow*) since it is more distant from the two other clusters; conversely, in Fig. 1e, the same overlaps with cluster 3 are subject to the strongest expansion when more overlaps are required ($\lambda = -0.5$).

Objective criterion (8) is used to guide the reallocation process in the usual Algorithm 1. Algorithm R_2-OKM thus follows the assignment heuristic detailed in Algorithm 2 but using the J_{R_2OKM} as objective criterion; finally, the sequential profile cluster updating is performed as follow:

$$p^*_{k,j} = \frac{\sum_i a_{i,k}\left[\left(\sum_l a_{i,l}\right)^{-2} p^i_{k,j} + \lambda \left(\sum_l a_{i,l}\right)^{-1} x_{i,j} \right]}{\sum_i a_{i,k}\left[\left(\sum_l a_{i,l}\right)^{-2} + \lambda \left(\sum_l a_{i,l}\right)^{-1} \right]}. \tag{9}$$

Let us notice that R_1-OKM and R_2-OKM are two generalizations of the c-means algorithm (and OKM); as for OKM, the strategies used for assignments and profiles updating ensure their convergence and have low cost in complexity with both algorithms in $O(TNK \log K)$ with T the number of iterations in the reallocation process. We also notice that the optimization process of the objective function through the different steps can lead to local optimum. This problem can be solved in practice by evaluating several clusterings with different initializations of profiles and maintaining only the one which gives the minimal value of the optimized criterion.

Table 1
Summary of the benchmarks used.

Dataset	Domain	Observ.	Dim.	Labels	Overlap
Iris	Botany	150	4	3	1
EachMovie	Video	75	3	3	1.14
Music emotion	Music	593	72	6	1.86
Scene	Image	2407	294	6	1.07
Yeast	Biology	2417	103	14	4.23

Table 2
Errors induced by additive and geometrical models on multi-label datasets. Symbol * marks preprocessed datasets (scaling and centering). Bold values highlight the best scores for each dataset.

Dataset	$\|\|X - AP^*\|\|_F^2$	$\|\|X - SP^*\|\|_F^2$
EachMovie	254.3	**128.4**
EachMovie*	**149.1**	150.4
Music emotion	2462127	**653903**
Music emotion*	36455.3	**36330.4**
Scene	23976.3	**19587.0**
Scene*	582831.0	**582399.2**
Yeast	**2284.1**	2293.5
Yeast*	**235476.3**	236393.9

4. Experiments

Clustering evaluation is known to be a difficult task in pattern recognition, mainly because of the vagueness of the definition of a "good clustering". Furthermore, validity measures traditionally used for clustering assessment are unsuitable for overlapping clustering. In addition to the visual assessment proposed in Fig. 1, we conduct in the following an external evaluation process using a new measure recently proposed and a suitable set of multi-label benchmarks.

4.1. Description of the datasets

We conduct experiments on different domains that motivate overlapping clustering researches[2]: video classification based on the users ratings (Eachmovie dataset), detection of emotion in music songs (Music emotion), clustering natural scene image (Scene) and genes (Yeast). The datasets are labelled with one or several labels for each data; they have been selected because of their diversity in

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

the application domain, the size (from 75 to 2417 data), the dimensionality (from 3 to 294 dimensions), the number of classes (from 3 to 14 labels) and the overlap rate (from 1 to 4.23 labels per data in average).

Table 1 summarizes the statistics of each dataset. Columns *Labels* and *Overlap* denote respectively the number of different proposed labels and the average number of labels assigned per data: the last one is given by $\frac{1}{N}\sum_i\sum_k a_{i,k}$.

Considering the four datasets with given multi-label references (EachMovie, MusicEmotion, Scene and Yeast), we study for each dataset wether a geometrical or an additive model is more likely to capture the expected overlaps. Let A be the true multi-label classification and X denoting the observations, we compute the optimal cluster profiles P^* considering either an additive or a geometrical model (cf. Section 2.2):

$$P^* = \begin{cases} (A^TA)^{-1}A^TX & \text{for the additive model} \\ (S^TS)^{-1}S^TX & \text{for the geometrical model} \end{cases}$$

Thus, for evaluating wether the cluster profile combination is more suitable with additive or geometrical combinations, we report in Table 2 a quantification of the sum of local errors $\|X - AP^*\|_F^2$ and $\|X - SP^*\|_F^2$ induced by each model respectively. As discussed in Section 2.2, this study reveals that EachMovie, Music Emotion and Scene are three datasets for which multi-labeling matches more with geometrical overlaps; a preprocessing (at least centering) on the data making additive models artificially suitable. Conversely, multi-labels in the Yeast dataset appear to fit with an additive modelling of the overlaps, thus testifying the success of such models for gene analysis [27,2].

4.2. Evaluation methodology

External evaluation of clusterings consists in quantifying the matching of a clustering with an expected classification given as a reference but not used during the clustering process. Most of the overlapping clustering methods have been assessed using standard F-measures considering precision and recall which are computed based on either labeling the clusters or counting the pairs of data linked by the clustering [2,6,5,20]. However, in the overlapping context, both methodologies present some limitations. The label based F-measure 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. On the other side, the pair-based F-measure solves the issue of labeling clusters but ignores the multiplicity of shared clusters and/ or shared labels between each linked pair of data. Amigó et al. [1] were the first to propose an extension from their BCubed metric that captures multiplicity precision and recall in order to deal with overlapping clusterings. As Suárez et al. [30] we use the extended F-BCubed measure for a fine-grained evaluation of the overlapping clusterings.

In our experiments we confront the new regulated methods (R_1-OKM and R_2-OKM) with five state-of-the-art methods: MOC and ALS[3] having the same underlying additive model but using different optimization strategies; OKM having geometrical model; traditional c-means which provides non-overlapping clusterings and a thresholded fuzzy-c-means. MOC, ALS, OKM and c-means are overlaps' parameter free, whereas R_1-OKM, R_2-OKM and the thresholded fuzzy-c-means require α, λ and a threshold σ respectively. For each parameter, several values are considered using the following strategies:

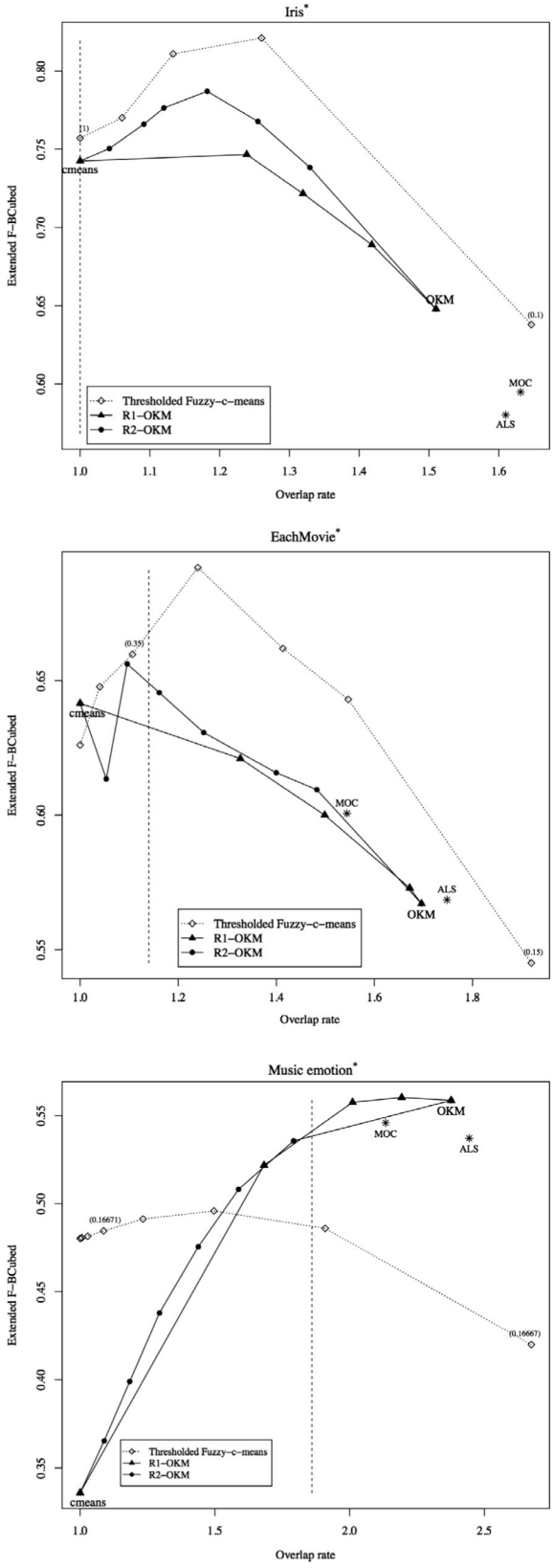

Fig. 2. Comparative positioning of the new overlap regulated models on datasets Iris, EachMovie and Music emotion. Symbol * marks preprocessed datasets (scaling and centering).

- α: we first detect by dichotomy on $[[0; 10]]$ the smallest value α_{min} leading to a non-overlapping clustering; then we test a hundred values uniformly distributed over $[0, \alpha_{min}]$ and keep

[3] Except for the Yeast dataset for which the number of clusters (14) proscribes to explore the combinatorial set of cluster combinations (ALS).

three salient values corresponding to strong jumps in the overlap rates obtained from two consecutive values of α.

- λ: we consider the seven following values $(0, 0.125, 0.25, 0.5, 1, 2, 5)$.
- σ: once the fuzzy membership matrix is learned, we search manually the minimum and maximum thresholds (σ_{min} and σ_{max}) leading to maximal and minimal overlap rates respectively. Then we test eleven values uniformly distributed over $[\sigma_{min}, \sigma_{max}]$.

As described, parameters α and λ are only considered for positive values because, on the datasets used, the regulations require mainly to restrain the overlaps with respect to the ones provided by OKM. Finally, let us mention the difficulty to tune manually the threshold σ as the number of clusters increases: some datasets require a 10^{-6} precision in order to capture an evolution between maximal and minimal overlap rates.

4.3. Results and discussion

Overlap rates have a strong influence on the matching measurement of the clusterings with respect to the expected classes. Thus, rather than providing tables of experimental results, we claim that a better way to actually understand the relative positioning of overlapping clustering methods consists in plotting their quality measures with respect to their overlap rates. Figs. 2 and 3 provide such a plotting for each of the five datasets with preprocessing of the data; each points on the figures are obtained by averages on ten runs of each algorithm with the same initial conditions (initial profile clusters) and a number of expected clusters equal to the number of labels. MOC, ALS and OKM being free of parameters controlling the overlap rate, only one solution is provided for each run leading to a single averaging point in the plots. Conversely, since R_1-OKM, R_2-OKM and the thresholded fuzzy-c-means allow the overlap regulation with a parameter to tune, we linked the consecutive points in order to represent the trends; the vertical dotted lines denote the expected overlap rates as given by the multi-labels datasets.

The analysis of the experimental results firstly shows the reliability of the extended F-Bcubed measure that is higher for clusterings whose overlap rate comes closer to the expected one on the whole. The main lesson we learned from the experimental results is the ability of the regulation principles to produce more fitting overlapping clusterings where the original model build clusterings with too large overlaps: on Iris, EachMovie, Music emotion and Scene, the original unregulated models provide highly overlapping clusterings, leading to degraded scores; the proposed regulations offer ways to reduce the overlaps while preserving the original overlapping clustering principle thus raising significantly the matching scores. Let us notice on Iris, EachMovie and Scene, the even better scores that R_2-OKM obtains with little more overlaps than expected: it can be explained by considering that overlaps also model a sort of indecisiveness on the assignment that is preferable to a wrong decision in the matching evaluation. Indeed we observed using R_2-OKM on Iris that all the imprecision comes from the two last classes (Versicolor and Virginica) both for crisp-clusterings ($\lambda \to +\infty$) and for clusterings with moderated overlaps (e.g. $\lambda = 0.5$); such overlaps are reliable and benefit the matching with the initial labeling.

When comparing the three parameterized methods we notice from the experimental results that, on the whole, the dispersal-based regulation (R_2) leads to better results than a regulation based on the number of assignments only (R_1). Regarding the thresholded

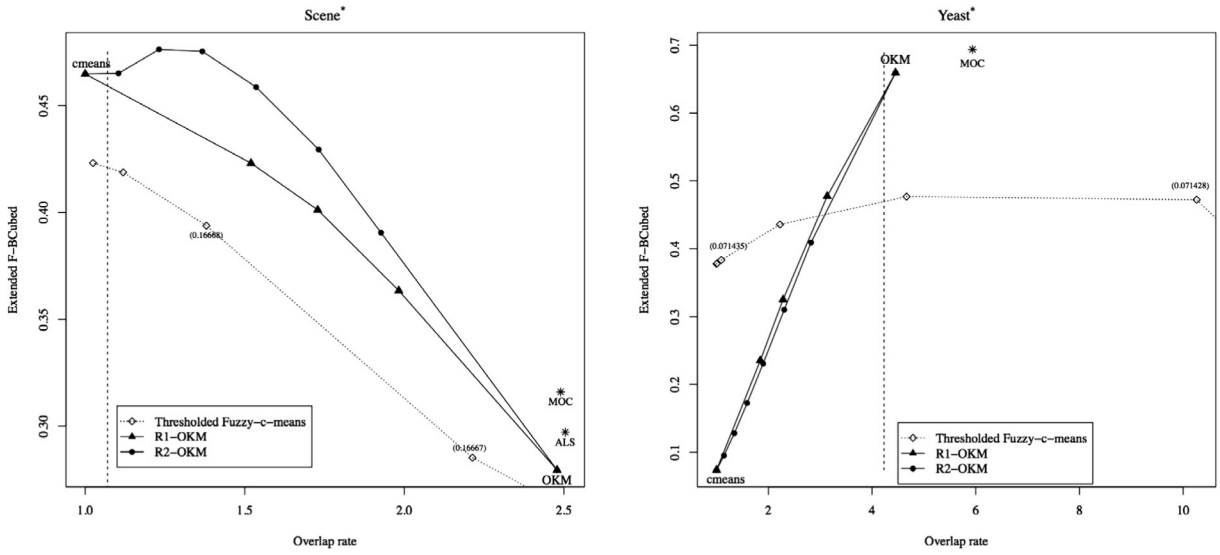

Fig. 3. Comparative positioning of the new overlap regulated models on datasets Scene and Yeast. Symbol * marks preprocessed datasets.

Table 3
Comparison of clustering models using the F-measure on preprocessed datasets (Iris, EachMovie, Music emotion, Scene and Yeast). Bold values highlight the best scores for each dataset.

Datasets	Iris*		Each Movie*		MusicEmotion*		Scene*		Yeast*	
	F measure	Overlap	F measure	Overlap	F measure	Overlap	F measure	Overlap	F measure	Overlap
c-Means	0.72 ± 0.04	1.00 ± 0.00	0.60 ± 0.06	1.00 ± 0.00	0.37 ± 0.01	1.00 ± 0.00	0.44 ± 0.02	1.00 ± 0.00	0.14 ± 0.00	1.00 ± 0.00
MOC	0.60 ± 0.02	1.63 ± 0.04	0.61 ± 0.03	1.54 ± 0.08	0.57 ± 0.02	2.13 ± 0.14	0.34 ± 0.03	2.49 ± 0.38	0.77 ± 0.08	5.94 ± 0.13
OKM	0.66 ± 0.05	1.51 ± 0.09	0.60 ± 0.02	1.70 ± 0.12	**0.63 ± 0.01**	2.38 ± 0.18	0.35 ± 0.01	2.48 ± 0.24	**0.81 ± 0.02**	4.45 ± 0.21
R_1-OKM	0.73 ± 0.04	1.24 ± 0.07	0.61 ± 0.03	1.33 ± 0.17	0.62 ± 0.02	2.19 ± 0.16	0.41 ± 0.02	1.52 ± 0.19	0.65 ± 0.12	3.13 ± 0.85
R_2-OKM	**0.76 ± 0.03**	1.18 ± 0.03	**0.62 ± 0.04**	1.10 ± 0.03	0.57 ± 0.01	1.79 ± 0.06	**0.45 ± 0.02**	1.23 ± 0.02	0.58 ± 0.01	2.82 ± 0.05

fuzzy-c-means, its achievement is actually limited since it outperforms R_1- and R_2-OKM only on datasets with few clusters (Iris and EachMovies with 3 clusters) and fail to build better overlaps when more clusters are expected (Music emotion, Scene and Yeast with 6, 6 and 14 clusters respectively); in addition, the thresholded fuzzy-c-means is much more sensitive to the threshold's tuning than R_1- and R_2-OKM are with their respective parameter α and λ, as illustrated by some threshold values mentioned in the figures.

Finally, to make easier a quantitative empirical comparison with previous studies, the different clustering models are compared using the usual F-measure and results are reported in **Table 3**. We notice that OKM-based models perform better than the non-overlapping c-means algorithm and also better than the MOC algorithm, despite the data preprocessing. In Addition, overlap regulations allow to improve the results obtained with OKM on three of the five datasets using both R_1 and R_2 regulation models[4] and with a strong advantage for the R_2-OKM method.

5. Conclusion

In this paper we focused on the overlapping clustering task, for which dedicated approaches provided unique clusterings, without regulation on the cluster overlaps until now. We introduced and formalized two regulated overlapping clustering models that generalizes the usual c-means approach to overlapping clustering. The two regulation principles aim to control the overlap shapes and sizes as regard to the number and the dispersal of the cluster concerned. We observed on real world datasets, the efficiency of the proposed principles and especially the ability of the second one (R_2-OKM) to build reliable overlaps with an easy tuning and whatever the requirement on the number of clusters.

We claim that the new principles introduced here are not limited to be applied with the geometrical OKM model and are easily transposable to additive models like ALS. One could also think of other ways of regulation as for example a single combination of the two present principles. But definitely, the major issue to address now concerns the use of the regulation principle in a new framework allowing to detect automatically a suitable regulation or to adjust such a regulation differently depending on the region space.

References

[1] E. Amigó, J. Gonzalo, J. Artiles, F. Verdejo, A comparison of extrinsic clustering evaluation metrics based on formal constraints, Inf. Retr. 12 (2009) 613.

[2] A. Banerjee, C. Krumpelman, S. Basu, R.J. Mooney, J. Ghosh, Model based overlapping clustering, in: International Conference on Knowledge Discovery and Data Mining, SciTePress, Chicago, USA, 2005, pp. 532–537.

[3] J. Baumes, M. Goldberg, M. Magdon-Ismail, Efficient identification of overlapping communities, in: Proceedings of the 2005 IEEE International Conference on Intelligence and Security Informatics, Springer-Verlag, 2005, pp. 27–36.

[4] P. Bertrand, M.F. Janowitz, The k-weak hierarchical representations: an extension of the indexed closed weak hierarchies, Discrete Appl. Math. 127 (2003) 199–220.

[5] F. Bonchi, A. Gionis, A. Ukkonen, Overlapping correlation clustering, in: 11th IEEE International Conference on Data Mining (ICDM), 2011, pp. 51–60.

[6] G. Cleuziou, An extended version of the k-means method for overlapping clustering, in: International Conference on Pattern Recognition ICPR, IEEE, Florida, USA, 2008, pp. 1–4.

[7] G. Cleuziou, L. Martin, C. Vrain, PoBOC: an overlapping clustering algorithm. Application to rule-based classification and textual data, in: R. López de Mántaras, L. Saitta (Eds.), Proceedings of the 16th European Conference on Artificial Intelligence, IOS Press, Valencia, Spain, 2004, pp. 440–444.

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

[9] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B 39 (1977) 1–38.

[10] D. Depril, I. Van Mechelen, B. Mirkin, Algorithms for additive clustering of rectangular data tables, Comput. Stat. Data Anal. 52 (2008) 4923–4938.

[11] E. Diday, Orders and overlapping clusters by pyramids, Technical Report 730, INRIA, France, 1987.

[12] M.R. Fellows, J. Guo, C. Komusiewicz, R. Niedermeier, J. Uhlmann, Graph-based data clustering with overlaps, Discrete Optim. 8 (2011) 2–17.

[13] Q. Fu, A. Banerjee, Multiplicative mixture models for overlapping clustering, in: Proceedings of the 8th IEEE International Conference on Data Mining, DC, USA, Washington, 2008, pp. 791–796.

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

[15] M. Goldberg, S. Kelley, M. Magdon-Ismail, K. Mertsalov, A. Wallace, Finding overlapping communities in social networks, in: 2010 IEEE Second International Conference on Social Computing (SocialCom), 2010, pp. 104–113.

[16] S. Gregory, An algorithm to find overlapping community structure in networks, in: Knowledge Discovery in Databases, Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, vol. 4702, 2007, pp. 91–102.

[17] S. Gregory, A fast algorithm to find overlapping communities in networks, in: Machine Learning and Knowledge Discovery in Databases, Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, vol. 5211, 2008, pp. 408–423.

[18] K. Heller, Z. Ghahramani, A nonparametric bayesian approach to modeling overlapping clusters, J. Mach. Learn. Res. 2 (2007) 187–194.

[19] P. Lingras, C. West, Interval set clustering of web users with rough k-means, J. Intell. Inf. Syst. 23 (2004) 5–16.

[20] H. Lu, Y. Hong, W.N. Street, F. Wang, H. Tong, Overlapping clustering with sparseness constraints, in: J. Vreeken, C. Ling, M.J. Zaki, A. Siebes, J.X. Yu, B. Goethals, G.I. Webb, X. Wu (Eds.), ICDM Workshops, IEEE Computer Society, 2012, pp. 486–494.

[21] M. Magdon-Ismail, J. Purnell, Ssde-cluster: Fast overlapping clustering of networks using sampled spectral distance embedding and gmms, in: Privacy, security, risk and trust (passat), 2011 IEEE Third International Conference on Social Computing (socialcom), 2011, pp. 756–759.

[22] M.H. Masson, T. Denoeux, Ecm: an evidential version of the fuzzy c-means algorithm, Pattern Recognit. 41 (2008) 1384–1397.

[23] B.G. Mirkin, Method of principal cluster analysis, Autom. Remote Control 48 (1987) 1379–1386.

[24] P. Pantel, D. Lin, Discovering word senses from text, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, Edmonton, Alberta, Canada, 2002, pp. 613–619.

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

[26] A. Pérez-Suárez, J.F. Martínez-Trinidad, J.A. Carrasco-Ochoa, J.E. Medina-Pagola, An algorithm based on density and compactness for dynamic overlapping clustering, Pattern Recognit. 46 (2013) 3040–3055.

[27] E. Segal, A. Battle, D. Koller, Decomposing gene expression into cellular processes, in: Pacific Symposium on Biocomputing, 2003, pp. 89–100.

[28] R.N. Shepard, P. Arabie, Additive clustering - representation of similarities as combinations of discrete overlapping properties, Psychol. Rev. 86 (1979) 87–123.

[29] C.G.M. Snoek, M. Worring, J.C. van Gemert, J.M. Geusebroek, A.W.M. Smeulders, The challenge problem for automated detection of 101 semantic concepts in multimedia, in: Proceedings of the 14th Annual ACM International Conference on Multimedia, ACM, New York, USA, 2006, pp. 421–430.

[30] A.P. Suárez, J.F.M. Trinidad, J.A. Carrasco-Ochoa, J.E. Medina-Pagola, An algorithm based on density and compactness for dynamic overlapping clustering, Pattern Recognit. 46 (2013) 3040–3055.

[31] L. Tang, H. Liu, Scalable learning of collective behavior based on sparse social dimensions, in: Proceedings of the 18th ACM conference on Information and knowledge management, 2009, pp. 1107–1116.

[32] Q. Wang, E. Fleury, Uncovering overlapping community structure, in: Communications in Computer and Information, Science, Complex Networks 116 (2011) 176–186.

[33] X. Wang, L. Tang, H. Gao, H. Liu, Discovering overlapping groups in social media, in: Proceedings of the 2010 IEEE International Conference on Data Mining, 2010, pp. 569–578.

[34] A. Wieczorkowska, P. Synak, Z. Ras, Multi-label classification of emotions in music, in: Intelligent Information Processing and Web Mining, Adv. Soft Comput. 35 (2006) 307–315.

[35] S. Zhang, R.S. Wang, X.S. Zhang, Identification of overlapping community structure in complex networks using fuzzy c-means clustering, Physica A 374 (2007) 483–490.

[4] For R_1-OKM and R_2-OKM, the scores reported in **Table 3** are the higher values observed considering the parameterization strategies mentioned in Section 4.2.