UNIVERSITE D'ORLEANS
Faculté des Sciences
LIFO

Laboratoire d'Informatique Fondamentale d'Orléans
4, rue Léonard de Vinci, BP 6759
F-45067 Orléans Cedex 2
FRANCE

Rapport de Recherche
www : http://www.univ-orleans.fr/SCIENCES/LIFO/

A Generalization of k-Means for Overlapping Clustering

Guillaume Cleuziou
Université d'Orléans, LIFO

Rapport N^o **RR-2007-15**

Abstract

This paper deals with overlapping clustering, a trade off between crisp and fuzzy clustering. It has been motivated by recent applications in various domains such as Information Retrieval or biology. We show that the problem of finding a suitable coverage of data by overlapping clusters is not a trivial task and we propose the algorithm OKM that generalizes the k-means algorithm combining a new objective criterion coupled with an optimization heuristic. Experimental results in the context of document clustering show that OKM first generates suitables overlaps between classes and then outperforms the overlapping clusters derived from fuzzy approaches (*e.g.* fuzzy-k-means).

1 Introduction

Clustering is a field of research belonging to both data analysis and machine learning major domains. Because new challenges appear permanently, new approaches have to be developed to deal with large amount of data, heterogeneous in nature (numerical, symbolic, spatial, etc.) and to produce several types of clustering schemes (crisp, overlapping or fuzzy partitions and hierarchies).

Many methodologies have been proposed in order to organize, to summarize or to simplify a dataset into a set of clusters such that data belonging to a same cluster are similar and data from different clusters are dissimilar. The clustering process is usually based on a proximity measure or, in a more general way, on the properties that data share. We can mention three major types of clustering processes according to the way they organize data: hierarchical, partitioning and mixture model methods [13, 3].

Most of the clustering methods have been developed in these frameworks in the last decades and allow a large amount of application fields. Nevertheless, some fields which led to recent attentions are still inefficiently processed. This is all the more true when the natural classes of data are neither disjoint nor fuzzy but clearly overlap. This situation occurs in important fields of applications such that Information Retrieval (several thematics for a single document), biological data (several metabolic functions for one gene). The present study aims at proposing a new theoretical framework coupled with an algorithmic solution for the task of structuring a dataset into suitable classes which overlap.

This paper is organized as follows: Section 2 describes related works which give only partial solutions to the overlapping clustering problem. Section 3 and Section 4 present respectively a new theoretical formalization and a first algorithmic solution (OKM) to this problem. The two last sections are dedicated to first experiments, variants (spherical-OKM) and discussions about the proposed approach with both theoretical and applied points of view.

2 Related works on overlapping clustering

A first way to produce overlapping classifications has been introduced by Jardine and Sibson [14]. They first proposed the k-ultrametrics which led more recently to the k-weak hierarchies [4], generalizing the previous pyramidal model introduced by Diday [9]. Even if these models are interesting because of the (visual) representation they produce, the overlapping schemes they allow are limited because in a pyramid each class can only overlaps with two other classes and a k-weak hierarchy have the following limitation: *"the intersection of (k+1) arbitrary clusters must be reduced to the intersection of some k of these clusters"*.

Another type of methods appears recently in order to deal with textual data particularly. CBC [16] and POBOC [7] for instance are based on heuristics and do not rely on criterion-based optimization. Furthermore, these methods are not free of thresholding problems. For example CBC requires four parameters, that are difficult to determine since the algorithm is very time consuming and proscribe multiple runs.

We can also mention an approach frequently used in practical situations: it consists in running well-known algorithms (k-means, fuzzy-k-means, EM, etc.) and modifying the result obtained to produce overlapping clusters. Modifications are performed by means of a threshold deciding whether an object belongs to a cluster or not, according to its proximity with the center of the cluster. This approach appears to be natural but it outlines two fundamental problems: first, the algorithm initially used does not take into account in the definition of the centers the fact that classes will overlap; secondly, the choice of a suitable (global) threshold, denoted above as the "thresholding problem", remains unsolved.

In case of using a partitioning algorithm first, the underlying hypothesis (h) assumed by the last approach is that *"we can reach a good overlapping scheme by extending a good partitioning scheme"*. We will show theoretically in the next section that this hypothesis is not satisfied generally; a new track so being necessary. In addition, we will show empirically that the second alternative which consists in the restriction of fuzzy classifications obtains worse results than an overlapping clustering model.

Finally, a (first) recent Model for Overlapping Clustering (MOC) has been proposed by Banerjee et al. [2]. It can be considered has a generalization of the EM algorithm. The main drawback for this kind of model-based approach is the multiple parameters to estimate (three matrices in MOC), making difficult large dataset processing. The present study follows an approach similar to MOC, and proposes a model able to deal with more data in a more simple way.

3 Theoretical framework

3.1 Definitions and notations

We introduce here the main definitions and notations used in the rest of the paper. We denote by X the set of data $\{x_1, x_2, \ldots, x_n\}$.

Definition 3.1 *Let $\mathcal{R}$=$\{R_1, \ldots, R_k\}$ be a set of classes over X, $\mathcal{R}$ forms a coverage of X if*

$$\forall x \in X,\ \exists R_j \in \mathcal{R} \mid x \in R_j$$

One can notice that definition 3.1 formalizes the notion of coverage in a broad sense since it allows nested and/or empty classes.

Definition 3.2 *Let $\mathcal{P}$=$\{P_1, \dots, P_k\}$ be a set of classes over X, $\mathcal{P}$ forms a partition of X if*

$$\forall x \in X,\ \exists! P_j \in \mathcal{P} \mid x \in P_j$$

We can notice that with the previous definitions, a partition is also a specific coverage.

In the following, we denote by $\mathcal{C}_{\mathcal{P}}$ the set of coverages obtained by extension of a partition $\mathcal{P}$. $\mathbf{P}$ and $\mathbf{P}_Q$ (resp. $\mathbf{R}$ and $\mathbf{R}_Q$) denote the set of all partitions (resp. coverages) and the set of all optimal partitions (resp. coverages) according to a criterion Q respectively.

3.2 Objective Criterion

The problem of finding an optimal partition according to an objective criterion Q is NP-hard because of the size of the search space. Indeed, even if the number k of classes is fixed there exists k^n ways to organize n data into k classes at the most. MacQueen was the first to propose a heuristic (k-means) to find a "locally" optimal solution starting from an arbitrary initial partition [15]. The success of the k-means algorithm is first due to its performance and also to its simple and intuitive underlying reasoning. Indeed, k-means is a linear time-complexity method and the process consists in iterating two steps: (1) computation of class centers (centers of gravity) and (2) assignment of each data to its nearest center. We can show that this process minimizes an intuitive criterion: the intra-class inertia (also called square error criterion) defined by

$$Q(\mathcal{P}) = \sum_{P_j \in \mathcal{P}} \sum_{x_i \in P_j} d^2(x_i, z_j)$$

with $Z = \{z_1, \dots z_k\}$ the centers of gravity for the classes $P_1, \dots, P_k$ respectively.

The final partition is only a local optimum because it depends on the initial partition considered. Main critics about the k-means method concern the problem of choosing a suitable initialisation and a suitable number of classes (k). Several solutions have been proposed, also applicable in our context [5, 17, 11].

The problem of finding a "good" coverage is also NP-hard since the search space is bigger again: there exists $2^{k.n}$ ways to organize n data into k overlapping classes at most. Furthermore the criterion $Q(.)$ used in the k-means approach is no longer suitable for coverages; we can actually show that $Q(\mathcal{R}) \geq Q(\mathcal{P})$ for all extension $\mathcal{R}$ of $\mathcal{P}$ ($\mathcal{R} \in \mathcal{C}_{\mathcal{P}}$). According to this objective criterion, a good coverage must be a partition. Then, a first step to address the overlapping clustering problem is to define a new objective criterion that allows to detect interesting coverages.

We propose a new objective criterion $\tilde{Q}$ that is an extension of the squarre error criterion Q. A partition $\mathcal{P}$ of X into k classes is actually defined by two sets of parameters :

- a set of binary membership values denoted by the matrix W ($k \times n$) with $w_{j,i} = 1$ if $x_i \in P_j$ (0 otherwise),

- a set of k class centers $Z = \{z_1, \ldots, z_k\}$.

The criterion Q measures for each data x_i, the error made when x_i is substituted by the only[1] center z_j such that $w_{j,i} = 1$. Being the representative of the class P_j, z_j is also the representative of x_i in $\mathcal{P}$ since $x_i \in P_j$.

In case of a coverage $\mathcal{R}$ of X into k overlapping classes, each data is then substituted by a set[2] of class centers $\{z_j | w_{j,i} = 1\}$. A natural way to define the representative of x_i in the coverage $\mathcal{R}$ is then to consider the center of gravity of this set. In the following this representative is called the "image" of x_i in $\mathcal{R}$ and is denoted by $\overline{x}_i$. Finaly the new objective criterion is defined by

$$\tilde{Q}(W, Z) = \sum_{x_i \in X} d^2(x_i, \overline{x_i})$$

with $\overline{x}_i$ the center of gravity of the set $\{z_j | x_i \in P_j\}$.

Let us notice that $\tilde{Q}(.)$ generalizes $Q(.,.)$, since in case of partitioning schemes, $\overline{x}_i$ is exactly the nearest center z_j.

3.3 Coverage vs. Extended Partition

Considering the new objective criterion $\tilde{Q}(.,.)$, we will show in this section that the hypothesis (h) mentioned in Section 2 is not satisfied. The hypothesis (h) assumes that

> *"we can reach a good overlapping scheme by extending a good partitioning scheme"*.

With the notations introduced in this section, (h) can be formalized as

$$(h) \quad \mathcal{P} \in \mathbf{P}_{\tilde{Q}} \Rightarrow \exists \mathcal{R} \in \mathcal{C}_{\mathcal{P}} \mid \mathcal{R} \in \mathbf{R}_{\tilde{Q}}$$

As an example, we consider $X = \{x_1, \ldots, x_6\}$ in $(\mathbb{R}^2, d)$ with d the Manhattan distance. Figure 1 provides the positions of the points and presents a partition $\mathcal{P}$ of X into 2 classes which minimizes the objective criterions $Q(\mathcal{P}) = \tilde{Q}(W, Z) = 12.0$ with W the membership matrix matching with Figure 1 and Z the centers of gravity of each class ($\mathcal{P} \in \mathbf{P}_{\tilde{Q}}$).

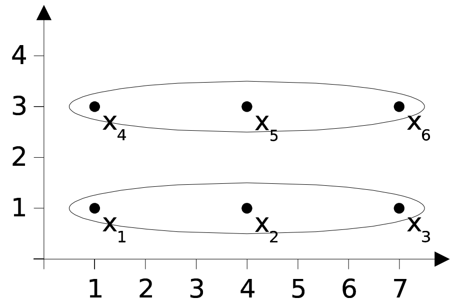

Figure 1: Example of a 2-class partition, optimal according to $\tilde{Q}$.

[1] In case of a partition, this center is unique because $\forall i, \;\sum_j w_{j,i} = 1$.

[2] In case of a coverage, $\forall i, \;\sum_j w_{j,i} \geq 1$.

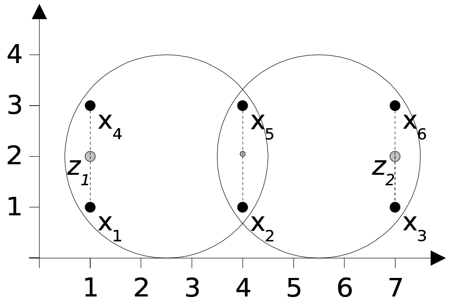

Figure 2: Example of the 2-class coverage, optimal according to $\tilde{Q}$.

An optimal coverage $\mathcal{R}^*$ on X according to $\tilde{Q}$ is defined by $\hat{W}$ and $\hat{Z}$ which minimize $\tilde{Q}(W, Z)$:

$$\mathcal{R}^* = (\hat{W}, \hat{Z}) \in \mathbf{R}_{\tilde{Q}} \Leftrightarrow \tilde{Q}(\hat{W}, \hat{Z}) \leq \tilde{Q}(W, Z) \ \forall (W, Z)$$

On this small size example we can observe only one optimal coverage $\mathcal{R}^*$ presented in Figure 2 (memberships, class centers and the center of gravity of the two centers). This coverage obtains a score $\tilde{Q}(\hat{W}, \hat{Z}) = 6.0$. By the way, we have shown with the previous counterexample that hypothesis (h) is not satisfied. Then, methods which consist in searching for an overlapping scheme starting from a partition reduce the search space with the risk that the considered subspace does not contains any good solution.

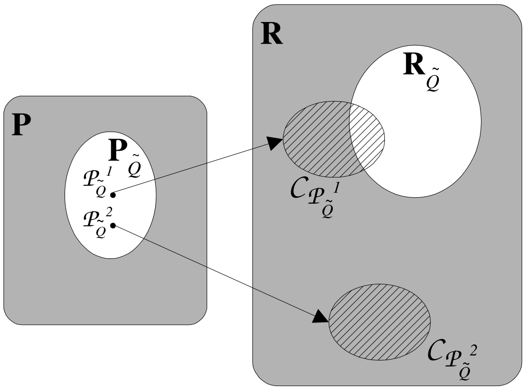
R
$\mathbf{R}_{\tilde{Q}}$
P
$\mathbf{P}_{\tilde{Q}}$
$\mathcal{P}^1_{\tilde{Q}}$
$\mathcal{P}^2_{\tilde{Q}}$
$\mathcal{C}_{\mathcal{P}^1_{\tilde{Q}}}$
$\mathcal{C}_{\mathcal{P}^2_{\tilde{Q}}}$

Figure 3: Partitions and coverages search spaces.

Figure 3 illustrates this phenomena, showing two situations: $\mathcal{P}^1_{\tilde{Q}}$ can reach an optimal coverage because some of its extensions are optimal and $\mathcal{P}^2_{\tilde{Q}}$ reaches only non-optimal coverages. In the following of this paper, we propose a search strategy in the global set of coverages to approximate optimal coverage by searching suitable parameters W and Z.

4 The algorithm OKM

We present in this section the algorithm OKM (*Overlapping k-means*) as a heuristic to approximate optimal coverages according to the generalized square error criterion $\tilde{Q}(.,.)$. The global process for OKM is similar to the k-means algorithm. A first initialisation step is followed by the iteration of two steps: computation of class centers and assignments, until a stable coverage is obtained (Figure 4).

```
Initialisation: $t$=0
  choose arbitrary $k$ centers $Z^t = \{z_1^t, z_2^t, \dots, z_k^t\}$ from $X$,
  For each $x_i \in X$: Assign($x_i$,$Z^t$) (build $W^t_{.,i}$),
  Build a first coverage $\mathcal{R}^t = (W^t, Z^t)$.
Do
    $t$=$t$+1
    • Update($Z^{t-1}$,$W^{t-1}$) (build $Z^t$),
    • For each $x_i \in X$: Assign($x_i$,$Z^t$) (build $W^t_{.,i}$),
While ( $W^t \neq W^{t-1}$ or $\tilde{Q}(W^{t-1}, Z^{t-1}) - \tilde{Q}(W^t, Z^t) < \epsilon$)
```

Figure 4: Pseudo-code of OKM.

The main differences with k-means concern the way to assign each data to one or several classes (multi-assignment) and the method used to update class centers. These two steps must ensure the decrease of $\tilde{Q}(.,.)$ in order to make the algorithm converge, and that the classes they lead to are of quality (classes of similar data).

```
Assign($x_i$,$Z$):
Initialisation :
  Let $z^*$ be the nearest center from $x_i$ in $Z$ ($\forall z_j \in Z,\ d(x_i, z^*) \leq d(x_i, z_j)$):
  $A = \{z^*\}$ (with $A$ the list of assignments for $x_i$),
  $Z = Z \setminus \{z^*\}$.
Do
  Let $\overline{x_i}^A$ denoting the center of gravity of $A$:
  Let $z^*$ be the nearest center from $x_i$ in $Z$,
    if $d(x_i, \overline{x_i}^{A \cup \{z^*\}}) < d(x_i, \overline{x_i}^A)$ then $A \leftarrow A \cup \{z^*\}$ and $Z = Z \setminus \{z^*\}$
While a new assignment is performed
Final decision:
  Let $A'$ be the old assignments for $x_i$,
    if $d(x_i, \overline{x_i}^A) < d(x_i, \overline{x_i}^{A'})$ then assign $x_i$ to the centers from $A$,
    else keep the old assignment $A'$.
```

Figure 5: Assignment process in OKM.

Let $Z = \{z_1, z_2, \ldots, z_k\}$ be a set of class centers and x_i a data to assign to one or several classes. The assignment process for x_i is described in Figure 5. It consists in scrolling through the list of centers from the nearest to the farthest, and assigning x_i while its image $\overline{x_i}$ is improved ($d(x_i, \overline{x_i})$ decreases). The new assignment is conserved only if it is better than the old one. The final decision enables to ensure that the objective criterion does not increase during the assignment steps.

Then, in the updating step, the new center z_j for a class R_j is defined in OKM by

$$z_{j,v} = \frac{1}{\displaystyle\sum_{x_i \in R_j} \frac{1}{\delta_i^2}} \times \sum_{x_i \in R_j} \frac{1}{\delta_i^2} . \hat{x_{iv}^j} \tag{1}$$

In (1), $z_{j,v}$ denotes the v^{th} component of the vector z_j, δ_i is the number of classes to which x_i belongs ($\delta_i = \sum_{j=1}^{k} w_{j,i}$) and $\hat{x_{iv}^j}$ denotes the v^{th} component of the center z_j "ideal" according to x_i, i.e. the center z_j such that $d(x_i, \overline{x_i}) = 0$. This last point is computed in the following way: $\hat{x_{iv}^j} = \delta_i . x_{i,v} - (\delta_i - 1) . \overline{x_{iv}}^{A \setminus \{z_j\}}$ where A is the set of classes to which x_i is assigned. We can propose a more intuitive definition of a new center z_j of a class R_j noting that z_j is the center of gravity of $\{(\hat{x_i}^j, \frac{1}{\delta^2}) | x_i \in R_j\}$ which is the set of "ideal" points where each point is weighted such that the more classes a data is member to, the less it impacts the new center position. We show below that new centers thus updated not only ensure that the objective criterion decreases but also enables to optimize (minimize) $\tilde{Q}(.,.)$.

Proof

Let X be a dataset into $(\mathbb{R}^p, d)$ where d is the euclidean distance, and $\mathcal{R}$ a coverage of X into k classes defined by memberships W and centers $Z = \{z_1, \ldots, z_k\}$ respectively.

Since OKM considers each class successively and separately during the updating step, it is sufficient to show that $\tilde{Q}(.,.)$ is minimized for each center z_j.

Decomposing X into two subsets according to whether data are members of R_j or not we obtain:

$$\tilde{Q}(W, Z) = \sum_{x_i \notin R_j} d^2(x_i, \overline{x_i}) + \sum_{x_i \in R_j} d^2(x_i, \overline{x_i})$$

For a data which is not member of R_j, its image is independent of z_j. Then, in the previous expression the left term is a constant (denoted as α in the following) relative to z_j.

We can rewrite $\overline{x_i}$ in the right term to bring a quadratic function relative to z_j:

$$\tilde{Q}(W, Z) = \alpha + \sum_{x_i \in R_j} \sum_{v=1}^{p} \left[x_{i,v} - \frac{1}{\delta_i} (\mathbf{z_{j,v}} + (\delta_i - 1) . \overline{x_{iv}}^{A \setminus \{z_j\}}) \right]^2$$

Then, $\tilde{Q}(W, Z)$ is minimized for a derivative equal to zero

$$\frac{\partial \tilde{Q}(W, Z)}{\partial z_j} = 0 \Leftrightarrow \sum_{x_i \in R_j} \frac{1}{\delta_i^2} \sum_{v=1}^{p} \left[z_{j,v} - \delta_i . x_{i,v} + (\delta_i - 1) . \overline{x_{iv}}^{A \setminus \{z_j\}} \right] = 0$$

On each component v the optimal center z_j is then defined by the following expression (equiv. to definition (1)):

$$z_{j,v} = \frac{1}{\displaystyle\sum_{x_i \in R_j} \frac{1}{\delta_i^2}} \times \sum_{x_i \in R_j} \frac{1}{\delta_i^2} \left[\delta_i . x_{i,v} - (\delta_i - 1) . \overline{x_{iv}}^{A \setminus \{z_j\}} \right]$$

□.

To close this presentation of the algorithm, we must notice that OKM can be seen as a generalization of the k-means algorithm. Indeed, if we decide to limit the assignments to only one ($\forall i,\ \delta_i = \sum_{j=1}^{k} w_{j,i} = 1$) we can notice that the new centers computed correspond to centers of gravity. An other important remark is to specify that OKM is non-deterministic because it depends on the initialisation (in the same way that k-means is non-deterministic) but also on the order the centers are computed during the updating step.

5 Experiments

Validation of classifications remains a difficult problem for unsupervised methodologies. Recent advances enable to distinguish three types of measures for cluster quality evaluation: relative, internal and external criteria [12]. The two first measures are mainly useful for comparing different partitions, different hierarchies and possibly different coverages, but they are clearly inefficient when the schemes that have to be compared are different in nature (e.g. partition vs. coverage).

Since we aim at evaluating the interest of overlapping schemes with respect to both crisp partitions and other overlapping schemes, we have decided to conduct experiments on datasets allowing external evaluations. External evaluation consists in comparing a set of classes obtained using a totally unsupervised process with a predefined organization on the same dataset. In this section we briefly observe the behavior of OKM on the well-known "Iris" dataset before to experiment OKM on the corpus "Reuters", that is one of the target practical domains (document clustering).

5.1 Iris dataset

The purpose of this preliminary experiment is to answer questions about OKM such as: the speed of convergence, the ability to retrieve the expected classes and to find interesting overlaps. The Iris dataset [10] is traditionally used as a test basis for a first evaluation. It is composed of 150 data in $\mathbb{R}^4$ tagged according to three non-overlapping classes (50 data per class). Using the euclidean distance and $k = 3$ we run OKM and k-means fifty times (with similar initializations) and we report the best result obtained according to $\tilde{Q}(.,.)$.
Figures 6 and 7 reports evolution of the values of the objective criterion during the iterations and confusion matrix respectively. We observe first that OKM and k-means have similar convergence speeds. We must notice that OKM (as for k-means), has a linear complexity on the size of the dataset with a complexity order of $O(t.n.k.\log k)$, with t the number of iterations, k the number of classes and n the number of data. We

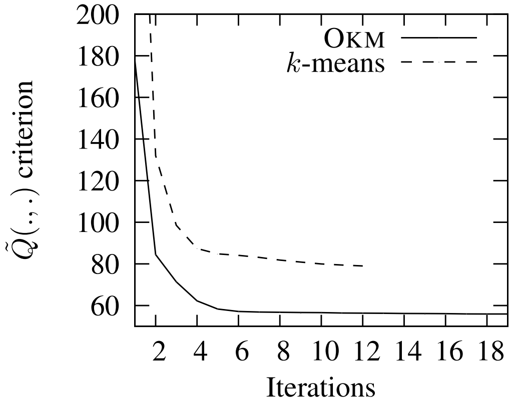

Figure 6: $\tilde{Q}(.,.)$ optimizations with OKM and k-means.

Tags\Clusters	1	2	3
Setosa			**50**/*50*
Versicolour	**26**/*3*	**50**/*47*	**9**/*0*
Virginica	**49**/*36*	**27**/*14*	

Figure 7: Confusion matrix for OKM (bold font) and k-means (italic).

can see on the confusion matrix that both methods are able to retrieve the expected classes according to the predefined tags.

To give the reader another way to assess overlaps between classes, we propose in Figure 8 a visualization of the Iris dataset with information about the classes obtained with OKM. This 3D-visualization comes from projection on the three first proper vectors (PCA). We then observe - on this projection - a natural cluster (top left) founded by OKM with few overlaps. Since the parameterization forced the algorithm to extract three classes ($k = 3$), OKM has extracted two other classes with many overlaps. Indeed, there is no natural separation between the set of remaining points.

5.2 Text Clustering with OKM

As mentioned in the introduction, Information Retrieval is one of the main target application for overlapping clustering. Indeed, a document can belong to several natural classes, for instance on the basis of the thematics it deals with.

5.2.1 Data preparation

We have performed the second experiment on the benchmark Reuters-21578[3] which is considered as one of the main dataset for text clustering or categorization experiments. This collection contains 21578 articles from press, written in english. Each article have one or several tags among a set of 114 predefined categories. From this full collection we have extracted a subset of 2739 documents such that:

[3] http://www.research.att.com/∼lewis/reuters21578.html

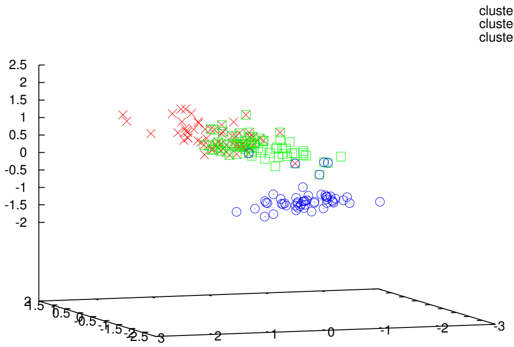

Figure 8: 3D-visualization of the classes obtained with OKM.

- at least one tag is proposed for the article,
- the body of the article is not empty,
- it appears in the "TEST" subset according to the splitting proposed by [1].

The indexing of documents is performed by using a process traditional in Information Retrieval. After tokenization and stemming [18], the p words with higher mutual information computed on the documents×words matrix are selected (tags are not used)[4]. Each document d_i is then indexed with a p-dimensional vector x_i such that each component v matches with a word w_v and $x_{i,v}$ is the frequency of the word w_v in document d_i. We use as proximity measure, the cosine similarity [19] which appeared to be the most efficient to compare two texts from a lexical analysis.

5.2.2 Spherical-OKM

We have previously proved that OKM converges towards a stable coverage into an euclidean space. Like for the k-means algorithm, a change is necessary to deal with the cosine similarity measure. This adaptation, denoted as spherical-k-means [8], consists in reasoning on the unit hypersphere in order to produce a partition maximizing the following criterion

$$Q_{\mathcal{S}}(\mathcal{P}) = \sum_{P_j \in \mathcal{P}} \sum_{x_i \in P_j} x_i^T . z_j$$

We notice that $Q_{\mathcal{S}}(.)$ can be seen as the opposite[5] of the square error criterion (to the normalization) with the cosine similarity measure.

To adapt OKM in the same way that k-means, we propose first to generalize $Q_{\mathcal{S}}(.)$ by

[4] In our experiment 423 words have been selected (multual information $\geq$ 0.1).

[5] Considering that similarity is the opposite of distance.

Nb. classes Algorithms	precision	recall	F-measure	size of the overlaps
k-means	0.61	0.08	0.14	1.00
Fuzzy-k-means (0.30)	0.58	0.08	0.14	1.00
Fuzzy-k-means (0.20)	0.52	0.10	0.17	1.29
Fuzzy-k-means (0.10)	0.55	0.15	0.23	1.51
OKM	0.53	0.18	0.26	1.87

Table 1: Comparison of results of OKM with k-means and Fuzzy-k-means methods for k=40.

$$\tilde{Q}_{\mathcal{S}}(W, Z) = \sum_{x_i \in X} x_i^T.\overline{x_i}$$

where $\overline{x_i}$ always denotes the image of x_i on the coverage $\mathcal{R}$ i.e. the center of gravity of the set of centers $\{z_j | w_{j,i} = 1\}$.

Then, the computation of new class centers must also be changed in OKM; we can show that the center which maximizes the objective criterion is, for each class, a simple (normalized) center of gravity on the set of data belonging to the class weighted contingently to their number of assignments $(\{(x_i, \frac{1}{\delta_i}) | x_i \in R_j\})$. This spherical variant of OKM can then be denoted as *spherical* -OKM.

5.2.3 Evaluation Framework and Results

On the Reuters dataset we use a relative criterion allowing us to compare: (1) the crisp classes obtained by a (spherical) k-means, (2) the overlapping classes obtained with a thresholded (spherical) fuzzy-k-means and (3) the overlapping classes resulting from (spherical) OKM. The relative criterion we use is a F-measure which combines precision and recall on the pairs of articles having a same tag in the predefined classification and the ones belonging to a same cluster in the partitions and coverages obtained.

$$\text{Precision} = \frac{\text{Number of Correctly Identified Linked Pairs}}{\text{Number of Identified Linked Pairs}}$$

$$\text{Recall} = \frac{\text{Number of Correctly Identified Linked Pairs}}{\text{Number of True Linked Pairs}}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Figures 9 and 10 and Table 1 report average values on fifty runs with different k. Each method have the same initialization for each run.

We observe first on Figure 9 that the choice of a suitable assignment threshold is difficult for fuzzy-k-means, because the sizes of overlaps is very dependant from the number of classes k. On the other hand we can see that OKM generates reasonable overlaps (from 1.5 to 2.0) with respect to the actual size of overlaps in the corpus (1.26), whatever the number k.

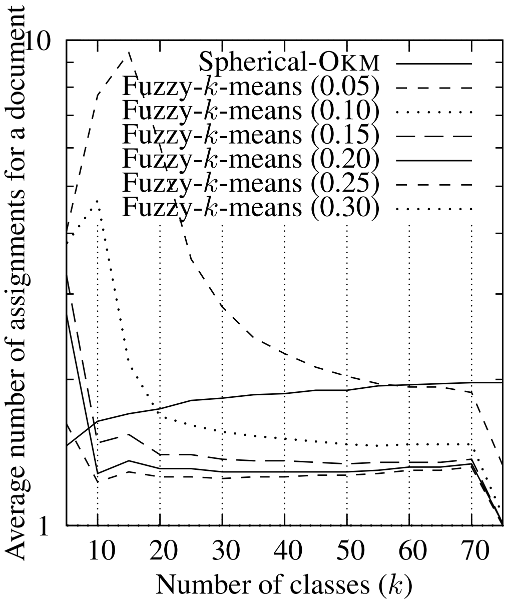

Figure 9: Average number of assignments (log scale) according to the threshold for fuzzy-k-means.

We selected the threshold 0.2 to be reported on the comparative study (Figure 10 and Table 1). We can then observe that scores decrease when k grows. This phenomena is natural since more classes implies less pairs of data, then a smaller recall. The main result to notice is that the relative criterion decreases faster for partitioning and Fuzzy-k-means overlapping schemes than for overlapping schemes obtained with OKM. Furthermore, an additional analysis show that the recall decreases in the same proportions whatever the clustering method. Then, the sensible better results observed with OKM are due to suitable overlaps which enable to associate more pairs of document with a good precision.

6 Conclusions and Further Works

The present study started from the following observation: clustering methods developed so far are not suitable to search an organization of data into overlapping clusters. However, this last type of classification scheme is becoming vital to deal with topical application domains such as information retrieval or bioinformatic.

We then proposed a new approach which aims at exploring the search space of possible coverages in order to retrieve a suitable organization into overlapping classes (or coverage). The approach presented is based first on the definition of an objective criterion which enables to evaluate overlapping schemes and then on the algorithm OKM as a heuristic to approach the optimal coverage according to the criterion. Both, criterion and algorithm must be seen as generalizations of the square error criterion and the k-means algorithm respectively.

Preliminary experiments showed a consistent behavior of the algorithm OKM (convergence and size of overlaps) and an ability to provide suitable overlaps especially for

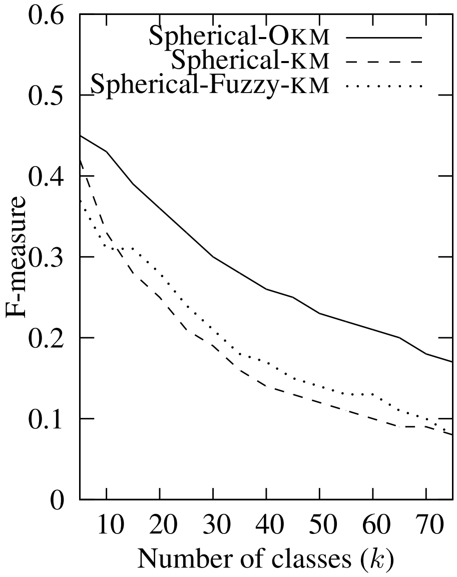

Figure 10: F-measure with different clustering methods.

text clustering which corresponds to one of the main target applications of this work. For this application, we proposed a variant (spherical-OKM).

We plan to progress about this study on two directions. First, we will proceed to other experiments: qualitative comparisons will be performed between OKM and other methods (like POBOC [7]) with datasets from other domains.

Secondly, it could be interesting to consider a (local) feature weighting for each class. This idea is based on [6]'s works and is meaningful in our framework since data should be assigned to each class on the basis of different features without inducing a rapprochement of the corresponding classes.

References

[1] C. Apté, F. Damerau, and S. M. Weiss. Automated learning of decision rules for text categorization. *ACM Trans. Inf. Syst.*, 12(3):233–251, 1994.

[2] A. Banerjee, C. Krumpelman, J. Ghosh, S. Basu, and R. J. Mooney. Model-based overlapping clustering. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 532–537, New York, NY, USA, 2005. ACM Press.

[3] P. Berkhin. Survey Of Clustering Data Mining Techniques. Technical report, Accrue Software, San Jose, CA, 2002.

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

[5] P. S. Bradley and U. M. Fayyad. Refining initial points for K-Means clustering. In *Proc. 15th International Conf. on Machine Learning*, pages 91–99. Morgan Kaufmann, San Francisco, CA, 1998.

[6] E. Y. Chan, W.-K. Ching, M. K. Ng, and J. Z. Huang. An optimization algorithm for clustering using weighted dissimilarity measures. *Pattern Recognition*, 37(5):943–952, 2004.

[7] G. Cleuziou, L. Martin, and C. Vrain. PoBOC: an Overlapping Clustering Algorithm. Application to Rule-Based Classification and Textual Data. In R. López de Mántaras and L. Saitta, IOS Press, editor, *Proceedings of the 16th European Conf. on Artificial Intelligence*, pages 440–444, Valencia, Spain, August 22-27 2004.

[8] I. Dhillon and D. Modha. Concept decomposition for large sparse text data using clustering. Technical report, IBM Almadan Research Center: RJ 10147, 1999.

[9] E. Diday. Orders and overlapping clusters by pyramids. Technical report, INRIA num.730, Rocquencourt 78150, France, 1987.

[10] C. B. D.J. Newman, S. Hettich and C. Merz. UCI repository of machine learning databases. University of California, Irvine, Dept. of Information and Computer Sciences, 1998.

[11] D.Pelleg and A. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 727–734, San Francisco, 2000. Morgan Kaufmann.

[12] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. Clustering Validity Checking Methods: Part II. *ACM SIGMOD*, 31(3):19–27, 2002.

[13] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[14] N. Jardine and R. Sibson. *Mathematical Taxonomy*. John Wiley and Sons Ltd, London, 1971.

[15] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical statistics and probability*, volume 1, pages 281–297, Berkeley, 1967. University of California Press.

[16] P. Pantel. Clustering by Committee. Ph.d. dissertation, Department of Computing Science, University of Alberta, 2003.

[17] J. Peña, J. Lozano, and P. Larrañaga. An empirical comparison of four initialization methods for the k-means algorithm. *Pattern Recognition Letters*, 20(50):1027–1040, 1999.

[18] M. F. Porter. An algorithm for suffix stripping. *Program 14*, pages 130–137, 1980.

[19] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1983.