An extended version of the k-means method for overlapping clustering

Guillaume Cleuziou
LIFO - University of Orléans
guillaume.cleuziou@univ-orleans.fr

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. We propose a new objective criterion and the associated algorithm OKM *that generalizes the k-means algorithm. Experiments show that overlapping clustering is a good alternative and indicate that* OKM *outperforms other existing methods.*

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: hierarchical, partitioning and mixture model methods [7, 2].

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 overlap. This situation occurs in important fields of applications such as Information Retrieval (several topics for one document) and biological data (several metabolic pathways for one gene). The present study proposes a 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 the few major works related to overlapping clustering. Section 3 presents the theoretical model and the algorithm OKM we propose. The next section is dedicated to experiments on real datasets before the conclusion.

2. Related works on overlapping clustering

A first way to produce overlapping classifications has been introduced by Jardine and Sibson [8]. They first proposed the k-ultrametrics which led more recently to the k-weak hierarchies [3], generalizing the previous pyramidal model introduced by Diday [5]. 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 approach frequently used in practical situations 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 cluster. This approach appears to be natural but it outlines two fundamental problems: first, the algorithm initially used aims at optimizing an objective function under constraints (hard or fuzzy assignments) that do not match with the expected clustering; secondly, the choice of a suitable (global) threshold, denoted above as the "thresholding problem", remains unsolved.

The method we propose is a center-based method that extends the k-means algorithm. We define a new objective function to minimize under constraints of multi-assignment. In a more general way, our approach explores the space coverages rather than the space partitions like k-means does. On this point of view, our approach is similar to the MOC algorithm recently proposed in [1] which is a kind of generalization of EM. We will show that MOC suffers from theoretical limits and algorithmic solutions which degrade its efficiency.

3. The OKM approach

3.1 Objective criterion

Given a set of data vectors $\mathcal{X}= \{x_i\}_{i=1}^n$ with $x_i \in \mathbb{R}^p$, the goal of the OKM algorithm (Overlapping k-means) is to find a k-way coverage $\{\pi_c\}_{c=1}^k$ of the date (where π_c represents the c^{th} cluster) such that the following objective is minimized:

$$\mathcal{J}(\{\pi_c\}_{c=1}^k) = \sum_{x_i \in \mathcal{X}} \|x_i - \phi(x_i)\|^2 \qquad (1)$$

Since $\{\pi_c\}_{c=1}^k$ is a coverage, each data x_i belongs at least to one cluster and the coverage is such that $\bigcup_{c=1}^k \pi_c = \mathcal{X}$. Thus, in (1) $\phi(x_i)$ denotes the "image" of x_i defined by combination of the prototypes (m_c) for the clusters x_i belongs to :

$$\phi(x_i) = \frac{\sum_{A_i} m_c}{|A_i|} \qquad (2)$$

In (2), A_i denotes the set of assignment for x_i : $\{m_c \mid x_i \in \pi_c\}$.

Let us notice that the new criterion $\mathcal{J}$ generalizes the least squared objective criterion used in k-means. Indeed, for single assignments $\phi(x_i)$ matches with the prototype of the only membership cluster for x_i.

3.2 Clustering algorithm

To minimize the objective $\mathcal{J}$ we propose a way to define cluster prototypes and to assign data to cluster in a traditional two-steps process.

The algorithm OKM starts with k prototypes $\{m_c^{(0)}\}_{c=1}^k$ drawn randomly in $\mathbb{R}^p$ or $\mathcal{X}$ and derives a first coverage $\{\pi_c^{(0)}\}_{c=1}^k$ by assigning data via a multi-assignment procedure we present below. Then OKM iterates the two following steps until a stopping criterion is reached :

1. the computation of new cluster prototypes $\{m_c^{(t+1)}\}_{c=1}^k$,

2. a multi-assignment procedure that leads to a new coverage $\{\pi_c^{(t+1)}\}_{c=1}^k$.

Like for the k-means algorithm the stopping criterion can be the convergence of the method, a maximum number of iterations or a threshold on the decreasing of the objective function. Figure 1 gives an overview of the algorithm OKM.

```
OKM($\mathcal{X}$,$t_{max}$,$\epsilon$)
Input: $\mathcal{X}$: a set of data vectors in $\mathbb{R}^p$, $t_{max}$: optional maximum number of iterations, $\epsilon$: optional threshold on the objective
output: $\{\pi_c\}_{c=1}^k$: final coverage of the points
1. Draw randomly $k$ initial cluster prototypes $\{m_c^{(0)}\}_{c=1}^k$ in $\mathbb{R}^p$ or $\mathcal{X}$.
2. For each $x_i \in \mathcal{X}$ compute the assignments $A_i^{(0)} = \text{ASSIGN}(x_i, \{m_c^{(0)}\}_{c=1}^k)$
   and derive the initial coverage $\{\pi_c^{(0)}\}_{c=1}^k$ such that
       $\pi_c^{(0)} = \{x_i | m_c^{(0)} \in A_i^{(0)}\}$
3. Set $t = 0$.
4. For each cluster $\pi_c^{(t)}$ successively, compute the new prototype
       $m_c^{(t+1)} = \text{PROTOTYPE}(\pi_c^{(t)})$
5. For each $x_i \in \mathcal{X}$ compute the assignments $A_i^{(t+1)} = \text{ASSIGN}(x_i, \{m_c^{(t+1)}\}_{c=1}^k, A_i^{(t)})$,
   and derive the new coverage $\{\pi_c^{(t+1)}\}_{c=1}^k$.
6. If not converged or $t_{max} > t$ or $\mathcal{J}(\{\pi_c^{(t)}\}) - \mathcal{J}(\{\pi_c^{(t+1)}\}) > \epsilon$, set $t = t+1$ and go to Step 4; Otherwise, stop and output final clusters $\{\pi_c^{(t+1)}\}_{c=1}^k$.
```

Figure 1. OKM **: Overlapping** k**-Means.**

3.3 Multi-assignment procedure

Given a set of k cluster prototypes, the assignment of each data to one or several clusters in a way such that the objective is minimized is not a trivial task. Indeed we cannot reasonably explore for every data points all the 2^k possibilities. Then we propose a heuristic (figure 2) that consists in scrolling through the list of prototypes from the nearest to the farthest, and assigning x_i while its image $\phi(x_i)$ is improved (in the sense of the squared euclidean norm). The new assignment is conserved only if it is better than the previous one (Step 3.), ensuring objective criterion decreasing.

```
ASSIGN($x_i$,$\{m_c\}_{c=1}^{k}$,$A_i^{old}$)
Input: $x_i$: data vector in $\mathbb{R}^p$, $\{m_1,\ldots,m_k\}$: set of $k$ cluster prototypes and $A_i^{old}$: optional multi-assignment
output: $A_i \subset \{m_1,\ldots,m_k\}$ a subset of cluster prototypes defining a multi-assignment for $x_i$
1. Set $A_i = \{m^*\}$ such that
       $m^* = \underset{\{m_c\}_{c=1}^{k}}{\text{argmin}} \|x_i - m_c\|^2$
   and compute $\phi(x_i)$ with assignment $A_i$.
2. Find the following nearest prototype
       $m' = \underset{\{m_c\}_{c=1}^{k} \backslash A_i}{\text{argmin}} \|x_i - m_c\|^2$
   and compute $\phi'(x_i)$ with assignment $A_i \cup \{m'\}$.
3. If $\|x_i - \phi'(x_i)\| < \|x_i - \phi(x_i)\|$ set $A_i \leftarrow \{m'\}$, set $\phi(x_i) = \phi'(x_i)$ and go to step 2;
   Otherwise, compute $\phi^{old}(x_i)$ with $A_i^{old}$
       if $\|x_i - \phi(x_i)\| \leq \|x_i - \phi^{old}(x_i)\|$ output $A_i$,
       else output $A_i^{old}$
```

Figure 2. Multi-assignment procedure.

3.4 Cluster prototypes calculation

Given a cluster π_h and a set of $k-1$ prototypes $\{m_c\}_{c=1}^{k} \setminus \{m_h\}$ the problem of finding m_h that minimizes $\mathcal{J}(\{\pi_c\}_{c=1}^{k})$ can be expressed as a convex optimization problem. The solution[1] is given by the following weighted average :

$$m_h^* = \frac{1}{\sum\limits_{x_i \in \pi_h} \alpha_i} \sum_{x_i \in \pi_h} \alpha_i . m_h^i \tag{3}$$

In expression (3), the weights α_i denotes the sharing of x_i among several clusters and is defined by $\alpha_i = 1/|A_i|^2$; the point m_h^i can be seen as the prototype m_h "ideal" for x_i, i.e. that would allow x_i to match with its image $\phi(x_i)$. Formally, m_h^i is given by

$$m_h^i = |A_i| . x_i - \sum_{m_c \in A_i \backslash \{m_h\}} m_c$$

To close the presentation of the algorithm, let us notice that the OKM iterative process generalizes the k-means algorithm in the sense that allowing only single assignment in OKM leads exactly to the traditional k-means procedure ($|A_i| = 1$, $m_h^i = x_i$ and $\alpha_i = 1$). Furthermore we notice that OKM inherits of the properties of k-means, it also has a linear complexity[2] on the size of $\mathcal{X}$ but it converges toward a local optima.

4 Experiments

We conducted experiments on real datasets from two different domains that motivate strongly overlapping clustering researches : Information Retrieval and Biology. F-measure is used as external criterion in order to compare different clustering methods; this criterion allows to measure the matching between the clusters obtained with OKM (and other clustering algorithms) and a predefined expected categorization (see [1] for details on this procedure).

We illustrate the usefulness of OKM for Information Retrieval applications by considering the text clustering task on the benchmark Reuters-21578[3]. The whole dataset contains 21578 articles belonging to one or several categories among a set of 114 topics. We built the three following subsets :

- Reuters-1 : contains the 1156 articles having at least one tag among the the set of 10 categories $\{$*gold, ipi,ship,yen, dlr, money-fx,acq,rice,grain and crude*$\}$.
- Reuters-2 : contains the 1308 articles having at least one tag among the the set of 10 categories $\{$*coffee, sugar, trade, rubber, earn, cpi, cotton, alum, bop and jobs*$\}$.
- Reuters-3 : contains the 333 articles having at least one tag among the the set of 10 categories $\{$*gnp, interest, veg-oil, oilseed, corn, nat-gas, carcass, livestock, wheat and soybean*$\}$.

For each dataset a feature set is extracted by selecting tokens that occurs into three articles at least. Reparameterization is then performed with LSA in order to provide a semantic indexing for each document.

We compare five clustering algorithms : (1) [KM] the k-means algorithm that produces crisp partitions, (2) [KM+] k-means + an additional soft-assignment step[4], (3) [FKM] the fuzzy-k-means algorithm + an additional soft-assignment step, (4) the [MOC] proposed by Banerjee et *al.* in [1] that builds overlapping clusters and (5) the [OKM] approach. Table 1 reports average F-measures on fifty runs with $k = 10$. For each run the five algorithms have the same initialization.

[1] Proof is not given in the paper for space convenience.

[2] The complexity order for OKM is $O(t.n.k \log k)$.

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

[4] Additional soft-assignment steps are performed via a threshold that is fixed empirically in order to obtain the best performances.

	Reuters-1	Reuters-2	Reuters-3
[KM]	0.500±0.0073	0.370±0.0015	0.332±0.0008
[KM+]	0.500±0.0073	0.370±0.0015	0.331±0.0000
[FKM]	0.447±0.0000	0.489±0.0000	0.338±0.0000
[MOC]	0.538±0.0033	0.694±0.0013	**0.347±0.0005**
[OKM]	**0.548±0.0003**	**0.761±0.0000**	0.339±0.0000

Table 1. Clustering on Reuters.

We first note that results are very similar for the algorithms [KM] and [KM+]; it illustrates the difficulty to determine a suitable threshold for additional assignments, particularly when the clustering is performed under constraints of single assignments. We also observe that dedicated overlapping approaches (MOC and OKM) obtain noticeable better results than traditional methods modified artificially via post assignments. Finally we notice that OKM outperforms MOC distinctly on the two first datasets and that OKM has a more stable behavior from one run to an other. The last remarks throw light on drawbacks of MOC which has recourse to k-means to performs the initialization step, among other things.

The second experiment is led on the domain of biology. On this domain comparative studies on real datasets are scarce because of the difficulty to work on reliable data (measures). Though we propose an experiment on the dataset of Gasch *et al.* [6] who characterized the genomic expression patterns of yeast genes in 15 different experimental conditions. Each of the 4373 gene is associated to one or several biological pathways among a set of 34 tags. We run 10 times the algorithms [KM], [MOC] and [OKM] on this dataset with $k = 34$ and observe the average results reported in table 2.

	Precision	Recall	F-measure
[KM]	**0.1658**	0.1234	0.1409
[MOC]	0.0562	**0.6936**	0.1040
[OKM]	0.0800	0.6046	**0.1413**

Table 2. Clustering on yeast genes.

This new experiment the previous one and show that performances of overlapping clustering can be explained by the additional assignments which involve a high recall in spite of the loss in the precision. The clusters produced by OKM have limited imprecision with respect to MOC, this results by a better F-measure.

5 Conclusion and perspectives

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. 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 clusters (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.

Experiments on real datasets showed a consistent behavior of the algorithm OKM and an ability to provide better overlapping clusters than other dedicated or extended algorithms.

We plan to purchase this study by considering a (local) feature weighting for each class. This idea has been led in [4] for traditional clustering but is meaningful in our framework since data should be assigned to each class on the basis of different features.

Acknowledgments

The research reported in this paper is partially funded by the French National Research Agency ANR (project GD2GS).

References

[1] 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.

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

[3] 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.

[4] 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.

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

[6] A. Gasch, P. Spellman, C. Kao, O. Carmel-Harel, M. Eisen, G. Storz, D. Botstein, and P. Brown. Genomic expression program in the response of yeast cells to environmental changes, 2000.

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

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