UNIVERSITÉ DE TUNIS
INSTITUT SUPÉRIEUR DE GESTION

MÉMOIRE DE MASTER RECHERCHE

SPÉCIALITÉ
SIENCES ET TECHNIQUES DE L'INFORMATIQUE DE DÉCISION

OPTION
INFORMATIQUE ET GESTION DE LA CONNAISSANCE

OVERLAP REGULATION FOR ADDITIVE OVERLAPPING CLUSTERING METHODS

MOHAMED ISMAIL MAIZA

SOUS LA DIRECTION DE:

NADIA ESSOUSSI MAITRE DE CONFÉRENCE, FSEG NABEUL
CHIHEB-EDDINE BEN N'CIR MAITRE ASSISTANT, ESEN MANOUBA

-

Laboratoire de Recherche Opérationnelle, de Décision et de Contrôle de processus (LARODEC)

ANNÉE UNIVERSITAIRE 2014 / 2015

Abstract

The identification of non disjoint clusters from unlabelled data sets is an important issue in clustering which is referred to overlapping clustering. This issue is based on the assumption that an object can belong to several clusters. The resolution of this issue solves many real problems which requires a non-disjoint partitioning of data such as image clustering where each image can contain several landscapes. The overlaps between clusters can have different sizes from one application to another. However, existing additive overlapping clustering methods produce a fixed size of overlap and do not give the user the possibility to regulate this size. Therefore, we propose in this work two additive overlapping clustering methods able to control the overlaps based on the user attempts. Experiments on artificial and real data set show the performance of the proposed method in adjusting the overlap size.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to Allah for giving me the strength, patience and wisdom to complete this work.

I would like to convey my warmest gratitude to my advisor **Dr. Nadia Essoussi** for giving me the opportunity to conduct my study in her research group and also for her precious advice, support, inspiration and continuous motivation.

My deepest gratitude also goes to my co-advisor **Dr. Chiheb-Eddine Ben N'Cir** for his guidance, generous contributions of knowledge, care, valuable comments and encouragement from the start until the end of my study.

I wish to express my sincere thanks to the **LARODEC** laboratory for providing great research conditions and all its members for their enthusiasm and determination.

I am particularly grateful to my **parents**, two elder **sisters** for their support and encouragement through my study .

Finally, I wish to thank my **friends** for being always standing by me through good and bad times.

Contents

Introduction **1**

1 Basic concepts of clustering and overlapping clustering **3**

1.1 Introduction 3

1.2 Basic concepts of clustering 3

1.2.1 Data representation 4

1.2.2 Proximity measures 6

1.2.3 Feature selection 9

1.2.4 Choosing the number of clusters K 10

1.3 Clustering methods 13

1.3.1 K-means 14

1.3.2 K-medoids 15

1.3.3 K-prototypes 15

1.3.4 Clustering schemes 16

1.4 Overlapping clustering 16

1.5 Overlapping partitional clustering methods 18

1.5.1 Uncertain membership based methods 18

1.5.2 Hard membership based methods 21

1.6 Conclusion 22

2 Geometrical and additive overlapping clustering methods 23

2.1 Introduction 23

2.2 Hard membership overlapping clustering 24

2.3 Additive overlapping clustering methods 25

2.3.1 Principal Cluster analysis 25

2.3.2 Alternating Least square ALS_{lf1} 26

2.3.3 Alternating least square ALS_{lf2} 28

2.3.4 Low dimensional additive overlapping clustering LAOC 31

2.4 Geometrical overlapping clustering methods 32

2.4.1 Overlapping K-means 33

2.4.2 Parametrized Overlapping K-means R_1OKM 34

2.4.3 Parametrized overlapping K-means R_2OKM 37

2.5 Conclusion 39

3 Overlap regulation for additive overlapping clustering methods 40

3.1 Introduction 40

3.2 Motivation 41

3.3 Alternating least square (ALS) 43

3.4 Additive overlapping clustering methods with overlap regulation 44

3.4.1 Proposed RALS1 method 44

3.4.2 Proposed RALS2 method 47

3.5 Step by step example for *RALS* 2 49

3.5.1 Execution of *RALS* 2 with $\alpha = -1.5$ 50

3.5.2 Execution of RALS2 with $\alpha = 1.5$ 53

3.6 Conclusion 57

4 Experimental study **58**

4.1 Introduction 58

4.2 Evaluation methodology 58

4.2.1 BCubed measures 59

4.3 Experiments on real and artificial multi-labelled datasets 61

4.3.1 Data sets description 61

4.3.2 Empirical results 62

4.4 Discussion 67

4.5 Conclusion 70

Conclusion **71**

References **73**

List of Figures

1.1 Unlabelled data (a) 4

1.2 Clustered data (b) 4

1.3 Manhattan distance and Chebychev distance 8

1.4 The role of feature selection in the clustering process 11

1.5 An example of overlapping clustering result 18

1.6 An example of uncertain memberships based clustering methods 19

2.1 Comparison of additive overlapping algorithms (Depril et al., 2008) 31

2.2 Example of voronoi cells induced by the R1OKM model (Ben NCir et al., 2014) . 36

2.3 Example of voronoi cells induced by the R2OKM model (Ben NCir et al., 2014) 38

2.4 Example of voronoi cells induced by the R1OKM and the R2OKM models (Ben NCir et al., 2014) 39

3.1 Expected clustering structure 41

3.2 Obtained clustering with ALS_{lf1} 42

3.3 Obtained clustering with ALS_{lf2} 42

3.4 $RALS2$ process with $\alpha = 1.5$ 56

3.5 $RALS2$ process with $\alpha = -1.5$ 57

4.1 Overlap size produced by every method on each data set 68

4.2 F-measure improvement with the use of overlap regulation on ALS_{lf2} 69

4.3 F-measure improvement with the use of overlap regulation on ALS_{lf2} 70

List of Tables

3.1 Example Dataset X 49

3.2 Clusters' profiles 50

3.3 Assignment matrix 52

3.4 Assignment matrix 56

4.1 Data sets description 62

4.2 Comparison of proposed methods with OKM, Parameterized ROKM, R_2OKM, ALS_{lf2} and ALS_{lf1} on the artificial dataset 63

4.3 Comparison of proposed methods with OKM, Parameterized ROKM, R_2OKM, ALS_{lf2} and ALS_{lf1} on the Iris dataset 63

4.4 Comparison of proposed methods with OKM, Parameterized ROKM, R_2OKM, ALS_{lf2} and ALS_{lf1} on the Emotions dataset 64

4.5 Comparison of proposed methods with OKM, Parameterized ROKM, R_2OKM, ALS_{lf2} and ALS_{lf1} on the EachMovie dataset 65

4.6 Comparison of proposed methods with OKM, Parameterized ROKM, R_2OKM, ALS_{lf2} and ALS_{lf1} on the Scene dataset 66

4.7 Comparison of proposed methods with OKM, Parameterized ROKM, R_2OKM, ALS_{lf2} and ALS_{lf1} on the Yeast dataset 67

List of Algorithms

1.1 The K-means Algorithm 15
1.2 The Partitionning Around Medoids Algorithm 15
2.1 The principal cluster analysis Pseudo-code (Depril et al., 2008) 27
2.2 Alternating least square lf1 Pseudo-code (Depril et al., 2008) 28
2.3 Alternating least square lf2 Pseudo-code (Depril et al., 2008) 30
2.4 Two-step reallocation process for overlapping clustering (Cleuziou, 2007) 34
2.5 Overlapping k-means assignment strategy (Cleuziou, 2007) 35
3.1 RALS1 pseudo-code 46
3.2 Assignment Strategy AS1 46
3.3 RALS2 pseudo-code 48
3.4 Assignment Strategy AS2 49

Introduction

Data mining aims to discover hidden patterns in large data bases. Clustering is one of data mining techniques which is able to discover the hidden patterns (Ganguly et al., 2008). This technique allows the detection of clusters from unlabelled data sets by organizing a given data set into heterogeneous clusters containing homogeneous observations. Clustering aims to fit structures in unlabelled data sets in the purpose to derive reduced representations of the data, to discover hidden links between observations, or to investigate the validity of a pre-existing group assignment. Clustering has been applied in several domains such as pattern recognition, image analysis, information retrieval, bio-informatics, market segmentation, sales segmentation, city planning.

While clustering aims to organize the data into disjoint clusters, many applications of clustering require non disjoint partitioning of data where data objects can be assigned to one or several groups at the same time. For example, in video classification a film can have different genres (McJones, 1997); In emotions detection, a piece of music can evoke different emotions at the same time (Trohidis, 2008) and in biology many genes need to belong to different clusters simultaneously (Battle et al., 2005).
In order to deal with this problem, many researches have been focusing on this problem known as overlapping clustering.

Overlapping clustering aims to detect the combination of clusters that best fit the data structure. Several overlapping clustering models were proposed such as hierarchical model (Diday, 1987) , Graph-based model (Fellows et al., 2011), correlational model (Bonchi et al., 2011) and partitional model (Banerjee et al., 2005). Our work is concerned with the partitional model. The

main concept of partitional overlapping clustering model consists in dividing the data into K groups by optimizing an objective function.

Partitional overlapping clustering can be defined on two families of methods which are additive and geometrical methods. Additive methods consider the overlap as the sum of the representatives of the related clusters such as PCL (MIRKIN, 1990) , ALS_{lf1} and ALS_{lf2} (Depril et al., 2008). Geometrical methods consider overlap as a geometrical logic in the data space such as OKM (Cleuziou, 2008), KOKM I (Ben N'Cir et al., 2015), KOKM II (Ben N'Cir et al., 2015). These methods generate fixed overlaps and do not fit with the data structure because the generated overlap do not usually match with the expected one. In fact, the existing partitional methods produce clusters with a large size of overlap and do not fit to the structure of the data. Ideally, the overlapping clustering method should produce an adjusted size of overlaps between clusters or give users the possibility to adjust this size.

Although, the regulation of overlap has been studied by (Ben NCir et al., 2014) for geometrical methods. We propose in this work to configure an overlap regulation principle for additive model of overlaps. We propose two additive methods with overlap regulation. The first one $RALS1$ extends ALS_{lf1} (Depril et al., 2008) while the second one $RALS2$ extends ALS_{lf2} (Depril et al., 2008). The proposed methods will be evaluated on artificial and real data sets.

This report is organized as follows : The first chapter reviews the different clustering basic concepts and recalls the necessary background for overlapping clustering. The second chapter presents the existing hard overlapping clustering methods. In the third chapter,we describe the proposed methods $RALS1$ and $RALS2$. The last chapter presents the different experiments performed on artificial and real overlapping data sets and shows the performance of the new methods. Finally, we conclude and we give some perspectives to this work.

Chapter 1

Basic concepts of clustering and overlapping clustering

1.1 Introduction

Clustering is an unsupervised machine learning tool which is usually performed as a preliminary step in data mining process. It aims to classify unlabelled data into groups that have meaning in the context of a particular problem.

In this chapter, we present the basic concepts of clustering in section 1.2 such as proximity measures such as feature extraction. In section 1.3 we describe the existing clustering methods. In section 1.4 we describe the related field of clustering which is overlapping clustering. Section 1.5 presents the partitional overlapping clustering methods. Finally, we describe in Section 1.4 an overview of existing overlapping clustering methods.

1.2 Basic concepts of clustering

One of the fundamental modes of understanding and learning is organizing data into sensible groupings. Cluster analysis is the study of grouping methods based on similarity measures without the use of category labels tagging objects with prior class labels. Thus, clustering can be

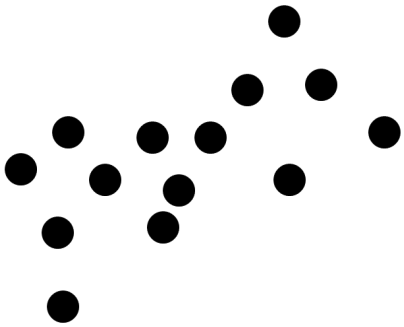


Figure 1.1: Unlabelled data (a)

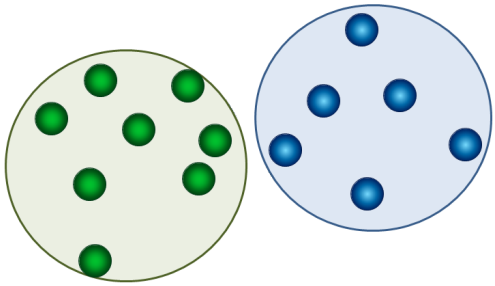


Figure 1.2: Clustered data (b)

seen as an unsupervised machine learning method where observations from a set of unlabelled data are assigned to k groups. The data assignment is based on the relationships between these observations. Observations in the same cluster have to be more similar than observations from different clusters. Clustering enables to fit structures in unlabelled datasets in order to derive a reduced representation of the full dataset. Clustering helps to discover new links between observations. Data Clustering analysis is used in many fields such as machine learning, pattern recognition, image analysis, information retrieval, bio-informatics, market segmentation, sales segmentation, city planning and many other fields.

Figure 1.1 shows how the unlabelled data is seen. Understanding a correct data structure is important. It can give much information about itself. Figure 1.2 demonstrates how the same data in figure 1.1 is classified after the clustering process, this step made the data more understandable and the shapes of the data have a meaning. Once the clustering process performed, the data is shown and interpreted in a different way. It gives clear and direct informations about itself.

1.2.1 Data representation

In order to fix the context and to make clear the prolific terminology that characterizes the data analysis domain, we assume that a data set consists of observations characterized with a set of attributes. These observations are also called data objects, instances or patterns. The attributes can also be called features, variables, dimensions or components and the clusters are called subsets, groups or classes. X denotes the data composed of n objects $X = \{x_1, x_2, \cdots, x_n\}$ where each object is characterized by m attributes $x_1 = \{x_{11}, x_{12}, \cdots, x_{nm}\}$. The data set X will be

represented as a matrix:

$$X = \begin{pmatrix} x_{11} & x_{12} \cdots & x_{1m} \\ x_{21} & x_{22} \cdots & x_{2m} \\ \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} \cdots & x_{nm} \end{pmatrix}.$$

The matrix X represents the attributes values for each object of the data set, i.e. the quantity in x_{11} represents the value of the attribute number one of the object number one.

Assignment representation

The assignment for each object is represented in an assignment vector that contains binary values indicating for each cluster if the object is assigned to that cluster or not. The assignment of the hole data set is represented in the following matrix A:

$$A = \begin{pmatrix} a_{11} & a_{12} \cdots & a_{1k} \\ a_{21} & a_{22} \cdots & a_{2k} \\ \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} \cdots & a_{nk} \end{pmatrix}.$$

The value $a_{52} = 0$ means that the fifth object is not assigned to the second cluster and the value $a_{51} = 1$ means that the fifth object of the data set is assigned to the cluster number one.
The matrix P represents the centroids of the clusters or the clusters' profiles. For each cluster, according to the clustering method used, the clusters' profiles will be computed taking in consideration each attribute and stocked into the following matrix P:

$$P = \begin{pmatrix} p_{11} & p_{12} \cdots & p_{1m} \\ \vdots & \vdots & \vdots \\ p_{k1} & p_{k2} \cdots & p_{km} \end{pmatrix}.$$

As an example, the value of p_{21} represents the profile of the cluster number two, accurately at the attribute number one.

1.2.2 Proximity measures

Proximity concept is acquiring more and more interest in the clustering applications since it defines how observations are interrelated. When the likeness between objects in the cluster is higher, the probability of a particular object to be assigned to a particular cluster is higher. Proximity measures functions are employed in clustering to compare objects with others in order to help to determine if the objects in question can be allocated to that cluster or not. Choosing the convenient similarity measure can be as major as the picking of the data representation, the number of clusters or the feature selection. Similarity measures take objects pairs as input and return numerical values that are as high as the objects are alike and as low as the objects are unlike. Dissimilarity measures return numerical values that are as low as the objects are alike and as high as the objects are unlike. Similarity measures quantify the strength of relationships between two items while dissimilarity measures quantify the degree of disorder between two items. A clustering result quality relies on the proximity measure used by the clustering analysis process. Thus, selecting a convenient measure for a specific problem is a tricky task while different measures are used for binary, numerical and several types of data.
The proximity function is raised to be a distance if it satisfies the following conditions (Ackermann, 2009) :

- Non-negativeness: $D(x, y) \geq 0 \forall x, y \in X$.
- Identity $D(x, y) = 0$ if and only if $x = y$.
- Symmetry: $D(x, y) = D(y, x) \forall x, y \in X$.

The function D(x,y) is said to be a metric if this additional condition is fulfilled :

- Triangle inequality: $D(x, y) \leqslant D(x, z) + D(z, y) \forall x, y, z \in X$.

The decision on the distance measure selection is performed according to the data type. There exist distance measures for numerical, binary and categorical data. However, data types represent a rich domain of variety and for each type of data, there exist a different quantifying methodology. We present in the following distance measures for numerical attributes.

In the case of the numerical data distance measures, the proximity information relies on a single quantity which is defined as a distance or a scalar product between two vectors of values, there are diverse distance measures for numerical data :

Euclidean distance

The most commonly used distance measure is the euclidean distance. It represents the regular humans perception about distances in the real world. It is also known as L2 Norm. The euclidean distance between two data points x and y is defined as follows:

$$D(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}. \tag{1.1}$$

where x_i and y_i represents respectively the ith dimension of x and y.
The Euclidean distance fulfils all the previously mentioned four terms and therefore it is a metric (Xu et al., 2005). The benefit of this distance is that formed clusters are invariant to changes in the feature space. This method has also disadvantages, if one of the input attributes has a relatively wide scale, then the other attributes will be dominated (Lesot et al., 2008). Therefore, this distance was improved by the employment of weights to control the relative influence of each attribute in the comparison. This leads to the weighted euclidean distance which is equivalent to a linear modification in the data before the use of the euclidean distance.

Manhattan distance

Manhattan distance or City Block distance between two data points is the sum of the absolute differences of their coordinates, defined as:

$$D(x, y) = \sum_{i}^{m} |x_i - y_i| \tag{1.2}$$

where x_i and y_i represent the ith dimension of x and y respectively.
The Manhattan distance is also called rectilinear distance or L_1 norm. The distance between two data objects measured alongside axes at straight angles. In another way, it represents the horizontal and vertical components simple sum (Black, 2006).

Chebychev distance

Chebychev distance is also named the maximum value distance. It calculates the absolute magnitude of the differences between coordinates of a pair of objects where the result is the higher

of their differences along any coordinate dimension (Cantrell, 2000) :

$$D(x, y) = max\,|x_i - y_i|\,. \tag{1.3}$$

Chebyshev distance can be convenient if the dissimilarity between data objects is given by differences in single dimensions instead of all the dimensions considered at a time.
Figure 1.3 illustrates the difference between the Manhattan distance and the chebychev distance.

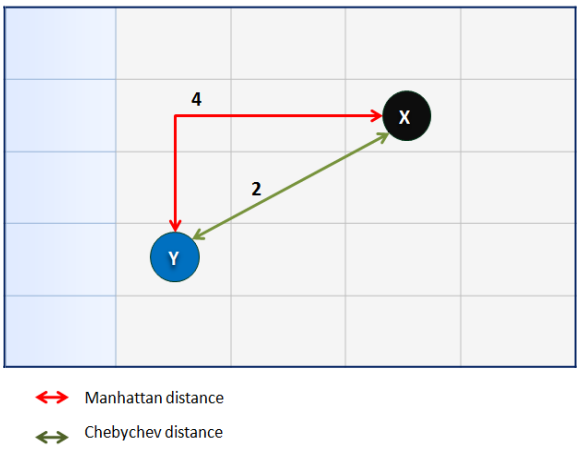


Figure 1.3: Manhattan distance and Chebychev distance

Minkowski distance

The definition of the minkowski distance of order p between two data points x_i and y_i is:

$$D(x, y) = \left(\sum_{i=1}^{m} |x_i - y_i|^p \right)^{\frac{1}{p}}. \tag{1.4}$$

Minkowski distance is considered as a generalisation of the Euclidean, the Manhattan and the Chbychev distances:

- p=2: Euclidean distance
- p=1: Manhattan distance
- $p \to \infty$: Chebychev distance

Mahalanobis distance

Mahalanobis distance represents the measure between two observations in the space defined by pertinent features. It takes into account various scales between data points and the correlation between them. Mahalanobis distance considers that variances in each orientation are unlike and it examines the covariances between variables to deal with the problem of scale and correlation mentioned with the euclidean distance. In the euclidean distance, a set of equidistant points from a determined position is a sphere that is stretched with the Mahalanobis distance to cover the respective scales of different data points and to consider the correlation between them. Thus, it helps to figure out abstract relations between data objects (De Maesschalck et al., 2000). The Mahalanobis distance is defined as:

$$D(x, y) = \sqrt{(x_i - y_i)^T S^{-1} (x_i - y_i)} \tag{1.5}$$

where x and y are two data objects, and S is the covariance matrix.
When S is a unit, the Mahalanobis distance becomes equivalent to Euclidean distance (Khachumov, 2012).

1.2.3 Feature selection

Applications requiring to cluster a large collection of data are various such as document clustering, gene clustering and clustering of earth science data. Appropriate features have to be sorted for the clustering operation, while inappropriate assumptions may bring faulty clusters. Therefore, variables have to be chosen regarding the conceptual and the practical considerations. Feature selection can be performed including only features that characterizes the data to be clustered (Jain, 2010). More precisely, the concept of feature selection consists in picking distinguishing variables by employing defined transformations. Feature selection aims to produce a useful features from the initial ones in order to make clustering analysis more efficient. Smart feature selection leads to powerful clustering characterised with a clear extraction of interpretations and an immunity to noise (Jain et al., 2000). When the space dimensionality is high enough, distances between closest points is no different of other points. Thereby, clustering algorithms based on distances may not be effective any more in high dimensional space. For this reason, dimension reduction is a major step in data preparation for cluster analysis in order to mine significant components from the original data.
Many dimension reduction techniques are commonly used :

- PCA (Principal Cluster Analysis): Consists in the creation of a linear combination of a set of vectors so as to better describe data variance (Oja, 1992)
- ICA (Independent Component Analysis): Relies on the use of statistical information with high order to detect the most statistically independent components from each others leading to distinguish independent signal sources from the mixed observation signals (Cherkassky & Mulier, 1998)
- MDS (Multidimensional scaling): Represents a non linear projection technique that adapts the initial high dimensional data into a low dimensional structure preserving the information of proximity (Duda et al., 2001)
- Isomap : Based on MDS : It estimates the Geodesic distance between a pair of points (Tenenbaum et al., 2000)
- LLE (Locally Linear Embedding): Uses a weight matrix describing relations between data points (Jain, 2010) .

In clustering, the concerned dataset with the analysis proceeds through the feature selection process in order to fit properly to the clustering algorithm as shown in Figure 1.4. The obtained features from the feature selection transformations are explored with the predefined clustering algorithm to give a set of clusters as a result. This latter is validated through clusters validation techniques. Once the resulting clustering is validated, a data structure is acquired.

1.2.4 Choosing the number of clusters K

The majority of clustering algorithms demand the predetermination of the number of the clusters before performing the clustering analysis process in order to insert it in the clustering process as a parameter. However, it is well known that this parameter influence directly the clustering process performance. The determining decision on the number of clusters is becoming increasingly important. This task is considered as one of the most difficult problems in data clustering. The reason is that it lacks standards and there are not much known techniques that helps in the definition or prediction of the number of clusters. There were some instinctive attempts aiming to deal with this problem. The distance between clusters in consecutive steps can be considered as a helpful guideline while the specialist can decide to stop the process when this distance overtakes a predetermined threshold. Another intuitive attempt to handle this problem consists

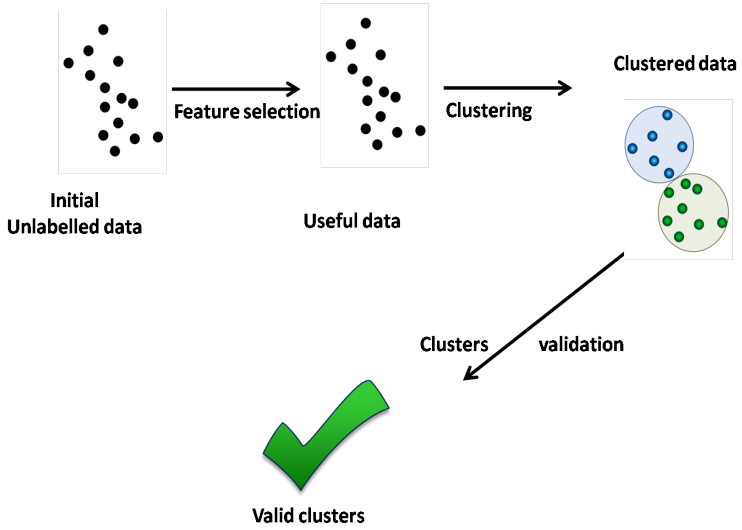


Figure 1.4: The role of feature selection in the clustering process

in calculating solutions of different numbers of clusters. The choice decision from those available solutions is performed in regard of a priori criteria, a practical opinion, or some theoretical basics. Intuitive relationships can also offer a nomination about the natural number of clusters. However, despite the ambiguity of the choice of the number of clusters problems, there exist three types of methods that aim to find a way to determine the optimal number of clusters.

Intra-cluster scatter based methods

Intra cluster based methods category consists in methods that take into account the within cluster depression-decay, which computes an error measure inside each cluster according to the within cluster depression method. The value of the error measure represents the pairwise sum distance for all objects in the cluster. This measure can be considered as a score indicating how are data objects dispersed from each others inside the cluster. The error measure decreases in function of the number of clusters and flattens at a certain number of clusters K. The value of K where the error measure flattens represents the optimal k for the concerned data clustering process. Another heuristic such as Proportional Reduction in Error (PRE) supposes that the intra-cluster scatter is a sum of the squared euclidean distances between the data objects and the cluster profile and then compares the result obtained with K clusters with the obtained result with K+1 clusters.

Intra and inter-cluster scatter based methods

The previously described methods ignores the inter-cluster distance. The intra and inter-cluster methods type takes into account simultaneously the inter-cluster scatter and the intra-cluster scatter. The main reasonable concept of the intra and inter-cluster scatter is to find a way to reduce the intra-cluster scatter while increasing at the same time the intra-cluster scatter. These methods determine a measure that computes the ratio between intra-cluster scatter and inter-cluster scatter. the validity index method considers the inter-cluster distance and the intra-cluster distance at the same time. This method relies on two measures which are the Mean Intra Cluster Distance (MICD) and the Inter-Cluster Minimum Distance (ICMD). It uses the behaviour of those two measures around the number of clusters K. MICD value is large in at least one cluster when the data is under-partitioned while it decreases sharply when the data shifts to over-partitioned. On the other hand, the ICMD represents a tiny value when the data is in the over-partitioned state. It uses the MICD and the ICMD measures lead to two new measures formalising the under-partition function and the over-partition function which helps to determine whether the data is under-partitioned or over-partitioned. The validity index is computed relying on these two measures and the optimal number of clusters is chosen with the smallest value of the validity index.

Based on criteria probabilistic methods

Using density-based methods, the determination of the appropriate number of cluster K becomes a submissive task offering the possibility of the use of clear probabilistic foundation. Criteria can be used in order to add new parameters that allows to obtain a better manner of fitting the data. There are several criteria that can be used such as the minimum message length (MML). Drichilet process (DP) and the minimum description description length. The minimum message length starts with a large number K of clusters and get clusters together while the criterion is decreasing. The Drichilet process is non parametric. It is used in the probabilistic models in order to generate a posterior distribution for the cluster's number.

Inspite the existance of these methods, The determining decision on the number of clusters is remains considered as one of the most difficult problems in data clustering. It still lacks standards and there are not much known techniques that helps in the definition or prediction of the number of clusters.

1.3 Clustering methods

Clustering aims to detect disjoint clusters from unlabeled data sets. Many researches have proposed several methods for clustering.

- Partitional methods: The concept of partitional clustering methods is to slice the dataset into k disjoint clusters. The acquired partitionning optimizes a certain criterion function which is iteratively optimized. The partitional clustering consists in creating an initial partionning, then a relocation technique is used to optimize the clustering by switching observations from one cluster to one other in order to optimize the criterion function. There exist three types of partitional clustering: The probabilistic approach, the centroid based approach and the medoid based approach. The probabilistic partitional clustering idea is that a mixture of populations forms the data, which leads to a draining enumeration of all the possible partitions in order to optimize the objective criterion. On the other hand, the main concept of the centroid based partitional methods is that the mean value of the assigned data to a cluster represents the partition. Analogously, the medoid based partitonal methods optimize their criterion function, the difference is that the partitions are represented with an observation assigned near the center. Examples of partional clustering methods are EM (Dempster et al., 1977), Partitionning Around Medoid (Kaufman & Rousseeuw, n.d.), PAM- CLARANS (Ng & Han, 2002) and K-means (?, ?)

- Hierarchical methods: Producing an hierarchy of clusters is the goal of the hierarchical clustering methods in order to exhibit the relations between the clusters. Hierarchical clustering methods consist in creating an hierarchical decomposition of the objects according to some criterions with many regroupmnet levels. Top-down and Bottom-up are the two types of the hierarchical clustering methods. Performing the top-down method consists in starting the clustering process with all the objects in one cluster. Then detaching the farthest objects into clusters until a termination condition holds, while using the bottom-up technique assigns each data object into one separate cluster, then merge the closest clusters into bigger ones. Examples of hierarchical clustering methods are BIRCH (T. Zhang et al., 1996) - Chameleon (Han et al., 2011)

- Density based methods: density based clustering methods aims to detect high density and low density regions among the dataset. High density region represent clusters and low density clusters indicate objects representing outliers. These methods avoid the use of

distances in the clustering process, but rely on the notion of density. When density in the neighbourhood exceeds a predefined threshold, the neighbourhood has to contain a fully minimum number of observations. The density based clustering concept makes it easily feasible to detect clusters with arbitrary shapes, while distance based clustering methods are able to detect only clusters with a spherical shape. Examples of density based clustering methods are DBSCAN (Density Based Spatial Clustering of Applications with Noise)(Ester et al., 1996) - MCLUST (Fraley & Raftery, 2003)

- Grid based methods: Grid based clustering methods are used for large data sets. A grid structure is created by transforming the space into a number of cells. Clusters' centres are identified in order to traverse the neighbourhood cells. Grid based clustering methods are characterised with a fast processing time in addition to the ability to treat highdimensional space. Examples of grid based clustering methods are WAVECLUSTER (Sheikholeslami et al., 2000) - BANG (Schikuta & Erhart, 1997)

1.3.1 K-means

K-means (MacQueen et al., 1967) is a clustering methods based on partitioning where each cluster is represented by its center. The center of each cluster is defined by the mean of the values of the observations assigned to that cluster. Partitioning a set of observations into K disjoint clusters depends on the choice of the number of clusters K and a distance measure. An initial partition is constructed randomly with the K not empty clusters. Each clusters' center is computed and each object is assigned to the cluster that have the closest center. Then each clusters' center is calculated again and every observation is assigned using the procedure until convergence. There are variant K-Means methods characterized by the way K is chosen. The distance metric and the methods calculating the means of the cluster. In the case where observations are represented with many attributes, normalization is required. The result depends on the initial clusters chosen. Therefore many trials are computed with different initializations. K-means helps in partitioning the data set into K disjoint clusters but allows an observation to be assigned to only one cluster as shown in Algorithm1.1.

Algorithm 1.1 The K-means Algorithm

1: Initialize: select random k data points as centres;
2: Assignment step: assign each data point to the cluster with the closest center;
3: Update step: recompute clusters centres ;
4: Repeat alternating step 2 and step 3 until there is no change in assignments;

1.3.2 K-medoids

K-medoids is a classical partitional clustering method used for categorical data. Medoids are used instead of means and the comparison is made by simple matching. The K-medoids method attempts to minimize the distance between points of an observation and points of the clusters' center where it should be assigned to. The highest frequency of a value of observations in a cluster defines the medoid which can also be defined as the object with the minimal average dissimilarity to all the other objects in the cluster. Meaning the most centrally located object in a cluster. The most used algorithm for K-medoids is PAM (Partitioning Around Medoids) (Kaufman & Rousseeuw, 1990) and is presented in Algorithm 1.2.

Algorithm 1.2 The Partitionning Around Medoids Algorithm

1: Initialize: select random k data points as medoids;
2: Assignment step: assign each data point to the cluster with the closest medoid;
3: Update step: select the data point with the lowest average dissimilarity to all other data points in the cluster as medoid;
4: Repeat alternating step 2 and step 3 until there is no assignment change;

1.3.3 K-prototypes

K-prototypes inherits the ideas of k-means and k-medoids. It is used to treat mixed types of data. A representative vector (prototype) for each cluster includes both of numerical and categorical data. For numerical attributes, the euclidean distance is used to compute distance between objects while for categorical attributes, the simple matching dissimilarity measure is employed. For example, the distance between pairs of the same category will be zero, and it will be one otherwise. Only one attribute value will be chosen to represent whole attribute in the clusters' center. Practically, the k-prototypes algorithm is more useful because in real world databases, frequently encountered objects are mixed type objects. A representative vector (prototype) for each cluster

i including both of numeric and categorical data:

$$C_i = (C_{i1}, C_{i2}, .., C_{im}). \tag{1.6}$$

where m represents the number of attributes C_{i1} = mean if the data is numeric and C_{i1} = medoid if the data is categorical.

1.3.4 Clustering schemes

Due to the fast evolution of clustering analysis, many methods have been proposed based on different techniques. Therefore clustering techniques can be classified in several categories according to the type of the clustering result obtained by each method. Three clustering schemes can be defined, namely, hard clustering, fuzzy clustering and the overlapping clustering. The main concept of hard clustering scheme is to slice the given data set into disjoint partitions representing the clusters where each object belongs to one and only one cluster. The second, fuzzy clustering scheme where the assignment of data points to partitions is represented relying on a degree of relationship between each data and each partition indicated by a membership function. This relationship degree is bounded by 0 and 1, 0 means no relation between the cluster and the object and 1 means that the object totally belongs to the cluster. The sum of the memberships degrees of the data object with all the clusters must be equal to one. This reasoning allows an observation to belong to one or several clusters but with different membership degrees. Anyway, the overlapping clustering scheme allows an object to be assigned to one or several clusters at the same time regardless of any degree of relationships. For instance, a movie can belong to the cluster of romantic movies and at the same time, it can belong to the cluster of action movies.

1.4 Overlapping clustering

Most of clustering applications in real world require observations to be assigned to more than one cluster. For example, in video classification, one movie can belong to multiple genres, in text mining, a document involve several themes. Therefore, the possibility that an observation can belong to one or different clusters is a crucial issue in many clustering applications. Overlapping clustering aims to find the combination of clusters that best fits the data. (Ben NCir et al., 2015) describe overlapping clustering methods as follows:

- **Hierarchical overlapping clustering methods:** The concept of the hierarchical methods is to limit the disharmony between the initial dissimilarities trough the examined dataset and the ones that the hierarchical structure generates. These methods are characterised with reduced overlaps since they don't consider all the possible combinations for each observation.

- **Graph-based overlapping clustering methods:** graph-Based overlapping clustering methods are generally used in the community detection context in complex networks. A graph represents the network with the vertices as data objects, and the links between the data objects are the edges. In order to cover the similarity graph, these methods employ a greedy heuristic that characterizes them with a high complexity level.

- **Generative overlapping clustering methods:** The assumption of overlapping clustering generative methods is that every data is the product of a blend of additive and multiplicative distributions. The probabilistic framework allows the treatment of any family of distributions. Although, these methods are not parametrizable, an overlap regulation can not take place by the expert.

- **Correlation and topological overlapping clustering methods:** The main concept of correlation based overlapping methods is presented as an optimization problem introducing an extension of the basic correlation clustering that allows overlaps to take place. Topological based overlapping clustering methods are perceived as an extension of the self organizing maps (SOM) (Cleuziou, 2013) by looking for a subset of neurons winners instead of looking for a single one. It grants to data object the possibility to been assigned to one or many neurons on the grid.

- **Partitional overlapping clustering methods:** Partitional overlapping clustering resides as the most commonly used. The main concept of these methods is to alter the resulting cluster from a basic method or to offer a new objective criterion in order to model the overlap. As follows, a description of these methods.

Figure 1.5 shows an example of overlapping clustering. Data objects are classified into three classes yellow, red and blue. Objects of the same color are close to each others and far from objects of different colors. Object in orange is assigned to the yellow and the red cluster at the same time. Object in green belongs to the yellow and the blue cluster simultaneously and likewise

with the purple coloured object. The black coloured object belongs to all the three clusters at the same time.

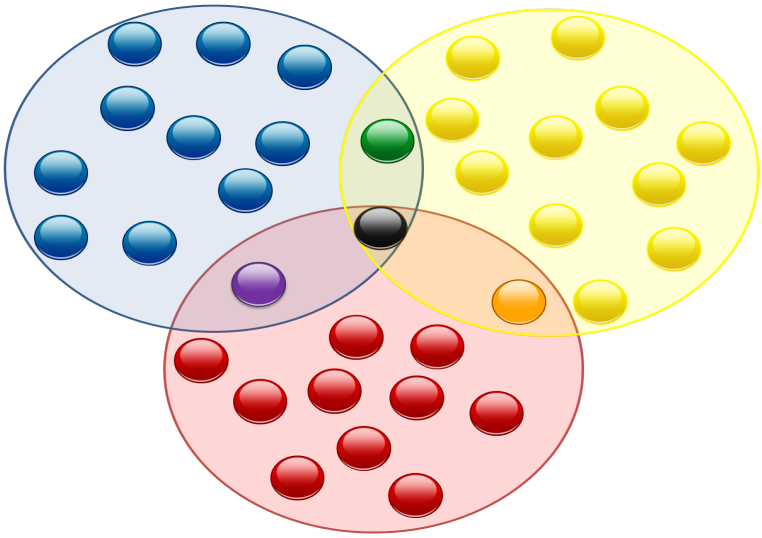


Figure 1.5: An example of overlapping clustering result

1.5 Overlapping partitional clustering methods

Many researches have been related to partitional clustering in order to obtain overlapping clusters. These researches led to two main categories of partitional clustering: the category of uncertain memberships and hard membership. Uncertain memberships represents the solutions modelling the clusters' memberships degree of each data using fuzzy, probabilistic or evidential frameworks as uncertainty function. Hard memberships represents the solutions modelling overlapping partitioning using a binary function.

1.5.1 Uncertain membership based methods

Extending uncertain methods results into overlapping clustering. This is the way uncertain memberships based methods in overlapping clustering works. Every uncertain membership methods requires a post-processing treatment to produce the resulting overlapping clusters. In the following, we describe the principal uncertain memberships-based methods producing non-disjoint

partitioning.

Figure 1.6 shows how the uncertain membership methods reach the overlapping assignment, firstly an assignment process takes place in order to accord to each observation a degree of relationship with each cluster, then a thresholding test gives birth to the final assignment overlapping matrix.

	Cluster 1	Cluster 2
Object 1	0.8	0.2
Object 2	0.3	0.7
Object 3	0.4	0.6

	Cluster 1	Cluster 2
Object 1	1	0
Object 2	0	1
Object 3	1	1

Threshold = 0.35

Figure 1.6: An example of uncertain memberships based clustering methods

Fuzzy c-means (FCM)

In the FCM (S. Zhang et al., 2007) , clusters are identified as fuzzy sets where an observation is affected to many clusters using a degree for memberships for all clusters. The fuzziness coefficient is in the interval [0,1] where 1 indicates a total membership to the cluster and 0 indicates no membership of the data to the cluster. FCM aims to minimize the following objective function:

$$J_{FCM}(A, P) = \sum_k \sum_i a_{ik} \left\| x_{ij-p_{kj}} \right\|^2 . \tag{1.7}$$

where A represents the fuzziness membership matrix indicating the relation of closeness of each data to every cluster:

$$a_{ik} \in [0 \cdots 1] \, \forall i, \forall k. \tag{1.8}$$

$$\sum_k a_{ik} = 1, \forall i \tag{1.9}$$

$$\beta > 1 \tag{1.10}$$

Parameter β controls the fuzziness of the memberships in a way where all memberships are set equals when β is set with a high value. When β tends to one, the fuzzy c-means works like a simple k-means algorithm. An alternating least square optimization of the two parameters A and P is performed iteratively in order to minimize the objective function. in each step the optimal fuzziness memberships matrix and the optimal clusters' representatives are computed as follows :

$$a_{ik}^{*} = \frac{1}{\sum_l (\frac{\|x_i - p_k\|^2}{\|x_i - p_l\|^2})^{\frac{1}{\beta-1}}} \tag{1.11}$$

$$p_k^{*} = \frac{\sum_i a_{ik}^{\beta} x_i}{\sum_i a_{ik}^{\beta}} \tag{1.12}$$

Extending fuzzy c-means to overlapping clustering can be achieved by using a threshold. Every observation's memberships degree that goes up to this threshold will be assigned to that clusters. When the threshold's value is high, some observations won't be assigned to any cluster if all it's memberships are lower than the threshold.

Possibilistic c-means (PCM)

Possibilistic c-means (Krishnapuram & Keller, 1993) identifies clusters as possibilistic sets allowing observations to belong to one or several clusters. A possibilistic type of memberships matrix was proposed to deal with the constrained condition in the fuzzy partition $\sigma a_{ik} = 1$.
PCM is based on the minimization of the following objective function:

$$J_{PCM}(A, P) = \sum_k \sum_i a_{ik}^{\beta} \|x_i - p_k\|^2 \tag{1.13}$$

under the constraints :

$$a_{ik} \in [0 \cdots 1] \, \forall i, \forall k \tag{1.14}$$

$$\sum_k (a_{ik}) = 1 \ \forall i \tag{1.15}$$

$$\beta > 1 \tag{1.16}$$

The minimization of the objective function is achieved iteratively using an alternating least square optimization of the two parameters A and P. Memberships and cluster's representatives are updated as follows :

$$a_{ik}^{*} = \frac{1}{1 + (\|x_i - p_k\|^2)^{\frac{1}{\beta-1}}} \tag{1.17}$$

$$p_k^* = \frac{\sum_i a_{ik}^{\beta} x_i}{\sum_i a_{ik}^{\beta}} \tag{1.18}$$

Possibilistic memberships can become overlapping without using a threshold. Because in the possibilistic theory, values doesn't matter, but the order of the values does. Overlapping memberships can be obtained by affecting the possibilistic memberships with higher values to 1, and the rest to 0.

Evidential c-means (ECM)

Defined as a general extension of fuzzy c-means (Masson & Denoeux, 2008) and possibilistic partitioning that leads to a credal partition. Exploring all the possible combinations of clusters and evaluating those combinations is the main idea of ECM using a mass of belief a_i within each possible combination. A is the credal partition matrix and P the matrix of clusters' representatives. Evidential c-means aims to minimize the following objective function :

$$J_{ECM}(A, P) = \sum_i \sum_{k/A_k \subseteq \Omega} p_k^{\alpha} a_{ik}^{\beta} \|x_i - \bar{p_k}\| \tag{1.19}$$

$$\sum_{k/A_k \subseteq \Omega} a_{ik} = 1 \forall i \in [1, \cdots, N] \tag{1.20}$$

Belief c-means (BCM)

Introduced as an extension of evidential c-means. Belief c-means (Liu et al., 2012) is developed within the belief function framework to improve the quality of the credal partiton by allocating only observations with a relative distance to the clusters combination. Belief c-means measures the distance between prototypes before the computation the mass of belief a_{ik} of each observation.

$$J_{BCM}(A, P) = \sum_{i=1}^{N} \sum_{k \backslash Ak \subseteq \Omega, |A_k|=1} a_{ik}^{\beta} \|x_i - m_k\|^2 + \sum_{i=1}^{N} \sum_{k \backslash Ak \subseteq \Omega, |A_k|>1} C_k^{\alpha} a_{ik}^{\beta} \widehat{d_{ik}^2} \tag{1.21}$$

where $\widehat{d_{ik}^2}$ estimates the clusters' prototypes inter-distance where x_i is assigned to.

Belief c-means, also named, hard credal partition, requires to an n data credal partitioning into 2^k possible combinations of k clusters. The main concept of belief c-means is to assign each observation to the set of clusters with the highest mass of belief

1.5.2 Hard membership based methods

Hard memberships based methods consist in generalizing the strict k-means in order to look for optimal overlapping clusters. Without requiring any post processing treatment, these methods produce hard overlapping clusters. There are two families of hard memberships based methods: Geometrical and Additive methods. Geometrical based methods describes overlaps as a barycentre on the related clusters' representatives. Minimizing the sum of distances between the average of clusters representatives and each observation such as the overlapping k-means which represents a generalisation of k-means, weighted overlapping k-means, kernel overlapping k-means with it's two variants KOKMI, a centroid based method and KOKMII a medoid based methods (Ben N'Cir et al., 2015), and the two parametrized overlapping k-means. On the other side, additive methods formalize overlaps as the addition of the representatives of the related clusters. Minimizing the sum of distances between the sum of cluster's representatives and each observation such as principal cluster analysis, the two Alternating least square ALS_{lf1} and ALS_{lf2}, and the low dimensional additive overlapping clustering. The partitional hard membership overlapping clustering methods are described in chapter 2.

1.6 Conclusion

In this chapter, we introduced a the basic concepts and required background for clustering such as clustering methods and similarity measures.We studied the overlapping clustering domain and we presented the different classes of overlapping clustering methods such as partitional overlapping clustering. We also studied uncertain membership overlapping clustering methods. The next chapter will be dedicated for an overview of hard membership overlapping clustering methods, such as geometrical and additive overlapping clustering methods.

Chapter 2

Geometrical and additive overlapping clustering methods

2.1 Introduction

Partitional hard membership overlapping clustering methods are the methods that allow the overlap to exist in a hard decision making manner. Unlike uncertain membership methods that requires post-processing treatments. Hard membership based methods generate as a direct result hard overlapping clusters. In the literatures, there are two types of hard membership overlapping clustering methods which are additive methods and geometrical methods. The additive methods family represents the methods adopting the hypothesis that the addition of the representatives of the concerned clusters is the overlap results. The geometrical methods consider the overlaps as a barycentre on the overlapping clusters. The geometrical methods family relies on a geometrical logic in the data space.
In section 2.2, we present a litterature review of hard membership overlapping clusteing methods. In section 2.3, we present Additive overlapping clustering methods while section 2.4 presents the geometrical overlapping clustering methods.

2.2 Hard membership overlapping clustering

(Shepard & Arabie, 1979) proposed the ADCLUS method for discovery and representation in similarity data. Later another method was proposed by (Arabie & Carroll, 1980) called MADCLUS that uses a penalty function on ADCLUS. INDCLUS (Carroll & Arabie, 1983) was proposed as a discrete model where subjects use the same subsets but patterns of weights vary with individual subjects. Then Mirkin proposed a one mode additive clustering method (Mirkin, 1987) using a set of vectors describing the profiles of clusters. In order to minimize the loss function of the Mirkin's model. Later, Principal Cluster Analyses PCL (MIRKIN, 1990) was proposed as a sequentially fitting algorithm.

According to the PCL algorithm, clusters are extracted one by one from the data set adopting the principal components extraction's steps. However, (Depril et al., 2008) found out some problems with the PCL in the recovering of the true models with a data set that produces overlapping clusters. In order to solve these problem, (Depril et al., 2008) proposed two new alternating least square methods ALS_{lf1} and ALS_{lf2} that are also based on the Mirkins model with the alternation of the estimation of the memberships matrix and the clusters profiles matrix. Later a new method based on Mirkins additive model by (Depril et al., 2012) low additive overlapping clustering method which deals with the problem of handling large data sets. The latter method performs the dimension reduction and clustering simultaneously.

The geometrical methods appeared with the emergence of the overlapping k-means method (Cleuziou, 2007) which is considered as a generalization of the simple K-means with a minimization of an objective criterion in order to produce overlapping clusters. Later, another two methods were proposed overlapping k-medoids and weighted overlapping k-means (Cleuziou, 2008) that represent respectively a generalization of the partitioning Around medoid method (PAM), and a generalization of both OKM and WOKM with weights of important features in the data in order to deal with the problem of the domination of certain attributes in the data. More lately, (Ben NCir et al., 2014) introduced two methods allowing the overlap regulation Parameterized ROKM and R2OKM based respectively on the number of assignments and the dispersal of cluster. Recently, (Ben N'Cir et al., 2015) the kernel overlapping k-means which is a generalization of the standard algorithm k-means that allows clustering in feature space. Another method emerged (Ben N'Cir et al., 2015) that performs the clustering process in high dimensional feature space where the objects are explicitly mapped relying on the positive definite kernels.

2.3 Additive overlapping clustering methods

Additive overlapping clustering methods consist in formalizing the overlaps as the addition of the representatives of the overlapping clusters. Minimizing the sum of distances separating the sum of the cluster's representatives and each observation. There are several additive methods for overlapping clustering such as principal cluster analysis, Alternating Least Square methods and the Low-dimensional Additive Overlapping Clustering method.

2.3.1 Principal Cluster analysis

Principal Cluster Analysis (MIRKIN, 1990) allows observations to belong to several clusters using the additive model. Calculating the sum of the clusters' representatives where the observation is assigned. Principal cluster analysis algorithm builds clusters memberships by starting with an empty cluster and sequentially assigning observations one by one regardless the memberships of other clusters in a greedy way until achieving the expected number of clusters k while minimizing the following objective criterion :

$$J_{PCL}(A, P) = \|X - AP\|_2^F = \sum_{x_i \in X} \left\| x_i - \sum_{k \in A_i} p_k \right\|^2 \tag{2.1}$$

Each time an observation is assigned to the cluster, a new value of J^c_{PCL} and a new cluster's representative valuep_k are calculated.

$$p_k^* = \frac{\sum_{x_i \in X} a_{ik} x_i}{\sum_{x_i \in X} a_{ik}} \tag{2.2}$$

PCL algorithm is characterised with a high computational complexity $O(2^{NK})$, for that reason, this method is rarely used in real world applications.

The clusters are explored on the dataset as follows:

$$\begin{cases} x_{ij}^m = x_{ij} - \sum_{k<m} a_{ik} p_{kj} & m = 2 \cdots k \\ x_{ij}^m = x_{ij} & m = 1 \end{cases} \tag{2.3}$$

where the $m - 1$ first clusters are removed from the data in order to obtain the mth remaining data x_{ij}^m. More precisely, the mth cluster is explored only from on the data x_{ij}^m regardless the

m-1 clusters. The prospection of the mth achieved while minimizing the following residual loss function :

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \left(x_{ij}^{m} - a_{im} p_{mj}\right)^{e} \tag{2.4}$$

and the conditionally optimal profiles values P_{mj} are calculated depending on the 0/1 cluster memberships $a_{im}(i = 1, \cdots, I)$ using the following function :

$$p_{mj} = \frac{\sum_{i-1}^{I} a_{im} x_{ij}^{m}}{\sum_{i-1}^{I} a_{im}}, for j = 1, \cdots, J. \tag{2.5}$$

where P_{mj} represents the average of the residual data of the objects in the cluster.
PCL algorithm starts with an empty cluster and inserts into it in a sequentially greedy way the object that gives the lowest loss value. Once the assignment of the object into the cluster performed, the cluster profile and new loss function are updated immediately and so on with every object until the loss function decreases no more. Thereafter, the product $a_{im} p_{mj}$ of the obtained membership vector and the cluster profile is subtracted from x_{ij}^{m} :

$$x_{ij}^{m+1} = x_{ij}^{m} - a_{im} p_{mj}. \tag{2.6}$$

The procedure is iterated with the new residual data x_{ij}^{m+1} , and the process continues until the number of cluster K is obtained.

2.3.2 Alternating Least square ALS_{lf1}

As an alternating least square type , ALS_{lf1} was proposed by (Depril et al., 2008) in order to minimize the loss function $lf(A) = \|X - A(A'A^{-1})A'X\|^2$ of the additive model of (Depril & Van Mechelen, 2005). This loss function is not separable, i.e. the two variables that the function aims to optimize are dependent of each others. The set of its elements are partitioned into series of subsets in order to update the memberships matrix. Each subsets is conditionally re-estimated upon the rest of subsets. The memberships matrix is updated row by row minimizing the following loss function :

$$J_{ALS}(A, P) = \sum_{i=1}^{N} \sum_{j=1}^{M} \left(x_{i,j} - \sum_{k=1}^{K} a_{i,k} p_{k,j} \right)^{2} \tag{2.7}$$

The following steps describe the process of the update of the memberships matrix. First of all, the algorithm treats the first row by searching for the binary combinations that brings the

Algorithm 2.1 The principal cluster analysis Pseudo-code (Depril et al., 2008)

Input: Data matrix X, number of clusters K.
m=1 ; $x_{ij}^m = x_{ij}$;
while $m \leq K$ **do**
  $a_m = (0, \cdots, 0)'$; length I
  $p_m = (0, \cdots, 0)'$; length J
  Calculate loss value
  do
    $l_{old}^2 = l^2$; $l_m^{old} = a^m$; $p_m^{old} = p^m$
    for all i such that $a_m(i) = 1$ **do**
      $a_m^i = a_m$; $a_m(i) = 1$
      Calculate p_m^i with memberships a_m^i
      Calculate loss value l_i^2 with a_m^i and p_m^i
    end for
    $i^* = argmin_i l_i^2$
    a_m and $p_m^{i^*}$
  while $l^2 < l_{old}^2$ and $a_m \neq (1, \cdots, 1)'$
  $a_m = a_m^{old}$; $p_m = p_m^{old}$;
  $x_{ij}^{m+1} = x_{ij}^m - a_m p_m$;
  $m = m + 1$
end while
Output : Matrix A with vectors a_m in in its columns.
Output : Matrix P with vectors p_m in its rows

minimal loss value from the set all the possible values given the rest of the rows of A_0. When the minimizing row is selected. It is inserted directly into the memberships matrix. This steps is performed repeatedly with the rest of the rows until all the rows are optimized. Then the algorithm compares the loss value obtained after the optimization process of the membership matrix with the loss value relative to initial memberships matrix. If there were no decrease in the loss function, the initial membership matrix will be kept, and the algorithm stops. In the other case where the loss function decrease with the new assignment matrix, the algorithm performs the same steps on the new memberships matrix in order to minimize the obtained loss. The algorithm stops when there is no further decreasing in the loss function. In this algorithm the optimal clusters profiles are recalculated instantly each time the membership matrix changes.

More precisely, when a new row takes place in the membership matrix, the new clusters profiles are updated before the search of another row in the same membership matrix in the beginning of the algorithm processing. The initial memberships matrix can be created from a Bernoulli distribution with $\pi = 0,5$ as parameter, or the conditionally optimal memberships matrix can be obtained up on K randomly drawn data points. As a result, we obtain a set of loss values each one lower than the one before this set of converging values. The algorithm is used with a multi-start strategy in order to avoid the local optimum trap.

Algorithm 2.2 Alternating least square lf1 Pseudo-code (Depril et al., 2008)

Input: Data matrix X, number of clusters K.
initialize $A_0 = A_1 : randomlydrawnoruserspecified$;
$P_0 = P_1 = (A_0^{'}A_0)^{-1}A_0^{'}X$;
$lf_2^0 = lf_2^1 = \|X - A_0P_0\|_F^2$;
do
$A_0 = A_1; P_0 = P_1; lf_1^0 = lf_1^1$
  for $i = 1; \cdots ; I$ **do**
    for $j = 0, \cdots , 2^K - 1 do$
      Convert into a binary (colomn) vector b of length K
      $A^j = A_1$;
      $A^j(i, :) = b; row\ i\ of\ A^j$
      Calculate loss $lf_{1;j} = lf(A^j)$
    end for
    $j^* = argmin_j lf_{1;j}$
    $A_1 = A^{j*}$;
  end for
  $P_1 = (A_1^{'}A_1)^{-1}A_1^{'}X$;
  $lf_1^1 = \|X - A_1P_1\|_F^2$;
while $lf_1^1 < lf_1^0$
Output : A, P.

2.3.3 Alternating least square ALS_{lf2}

Alternating least square algorithm (Depril et al., 2008) consists in estimating the conditionally optimal representatives P upon A. Then, it estimates the conditionally optimal memberships A

upon P, the conditional estimation of P and A will be repeated until convergence.

$$J_{ALS}(A,P) = \|X - AP\|_F^2 = \sum_{x_i \in X} \left\| x_i - \sum_{c \in A_i} p_k \right\|^2 \tag{2.8}$$

In order to estimate the optimal memberships, ALS_{lf2} estimates optimal memberships for each observation x_i by considering all possible combinations leading to decrease the objective criterion. This procedure is repeated for each observation x_i. Objective criterion's value is computed with the new memberships A_1 and compared with the value of the objective criterion computed with the previous memberships. If the new value of the objective criterion decreases, the new memberships are considered. Using theALS_{lf1}, the conditionally optimal representatives C are recalculated straight away for each memberships update. The optimal representatives are computed as follows :

$$P^* = (A^{'}A)^{-1}A^{'}X. \tag{2.9}$$

Alternating least square algorithm is a highly time consuming algorithm for the reason that it explores all the possible 2^k cluster's combinations for each observation. It is also based on Mirkins model (ALS type (MIRKIN, 1990)) in order to optimize the loss function :

$$lf(A,P) = \|X - AP\|_F^2 = \sum_{i=1}^{I} \sum_{J}^{j=1} \left(x_{ij} - \sum_{K=1}^{K} a_{ik} p_{kj} \right). \tag{2.10}$$

This algorithm alternates the membership matrix A and the profile matrix P in sequential manner, conditionally optimal upon the other. ALS algorithm starts with drawing randomly an initial membership matrix, then estimates the conditionally optimal profile matrix upon the initial memberships matrix using the Mirkin's model least squares multiple estimator:

$$lf(A,P) = \sum_j \left(x_{1j} - \sum_{k=1}^{K} a_{1k} p_{kj} \right)^2 + \cdots + \sum_j \left(x_{Ij} - \sum_{k=1}^{K} a_{Ik} p_{kj} \right)^2. \tag{2.11}$$

Relying on the freshly computed P_0 , the algorithm estimates a new membership matrix A_1 conditionally optimal upon computed P_0. For this calculation, a separability property of loss function lf2 is used (Chaturvedi & Carroll, 1994) and respected as follows:

$$lf(A,P) = \|X - AP\|_F^2 = \sum_{i=1}^{N} \sum_{j=1}^{M} \left(x_{i,j} - a_{i,k} p_{k,j} \right)^2. \tag{2.12}$$

This step leads to separate the intervention of ith row of A_1 to the loss function from the interventions of the other rows. As an outcome, the estimation of A_1 can be accomplished row by row

which helps to reduce the evolution work from the full 2^{IK} membership matrices to a $I2^K$ possible membership rows. Then a new matrix of profiles is estimated using the Mirkins one model clustering least square multiple regression estimator Equation 2.3.3. The process is repeated until there is no further decrease in the loss function. This process of alternations brings out a set of estimations that are conditionally optimal in a consecutive manner. It provides a row of loss values that does not increase and which is bounded with zero. The set of loss values converges anyway. The algorithms stops when the maximum number of iterations is reached, when the gap between the newly obtained loss value and the prior loss value is lower than the minimum optimization parameter or when there is no further decrease in the loss function.
In order to generate the initial membership matrix, two strategies can be performed. The main concept of the first one consist in taking the conditionally optimal memberships upon k randomly drown data points. the main concept of the second one consists in defining the entries of A_0 an independent and identical distributed from a Bernoulli distribution with $\pi = 0,5$ as a parameter. Usually, the alternating least square algorithm provides a local optimum because the resulting clustering is strongly related with the seeded Matrix A_0. For this reason, the ALS_{lf2} is generally employed with a multi-start strategy and the lower loss value will be selected as a final result of the method.
Figure 2.1 shows the behaviour of the PCL, ALS_{lf1} and the ALS_{lf2} with different overlap sizes.

Algorithm 2.3 Alternating least square lf2 Pseudo-code (Depril et al., 2008)

Input: Data matrix X, number of clusters K.
initialize $A_0 = A_1$: randomly drawn or user specified;
$P_0 = P_1 = (A_0^{'}A_0)^{-1}A_0^{'}X$;
$lf_2^0 = lf_2^1 = \|X - A_0P_0\|_F^2$;
do
    $A_0 = A_1; P_0 = P_1; lf_2^0 = lf_2^1$
    for $i = 1; \cdots ; I$ **do**
        Estimate row i of A_1 by minimizing enumeratively
        $\sum_j (x_{ij} - \sum_{k=1}^{K} a_{ik}p_{kj}^0)^2$
        over all possible 0/1 combinations for $a_{ik}, k = 1 \cdots , K$.
    end for
    $P_1 = (A_1^{'}A_1)^{-1}A_1^{'}X$;
    $lf_2^1 = \|X - A_1P_1\|_F^2$;
while $lf_2^1 < lf_2^0$
Output : A, P.

the ALS_{lf1} and the ALS_{lf2} were tested with both a multi-start procedure and with the data based starting strategy. In high overlap amounts, ALS_{lf1} and ALS_{lf2} used with a randomized multi-start procedure, manifest better performance than the ALS_{lf1} and ALS_{lf2} with the data based starting strategy. These latter show a better performance when the overlap size is low (Depril et al., 2008).

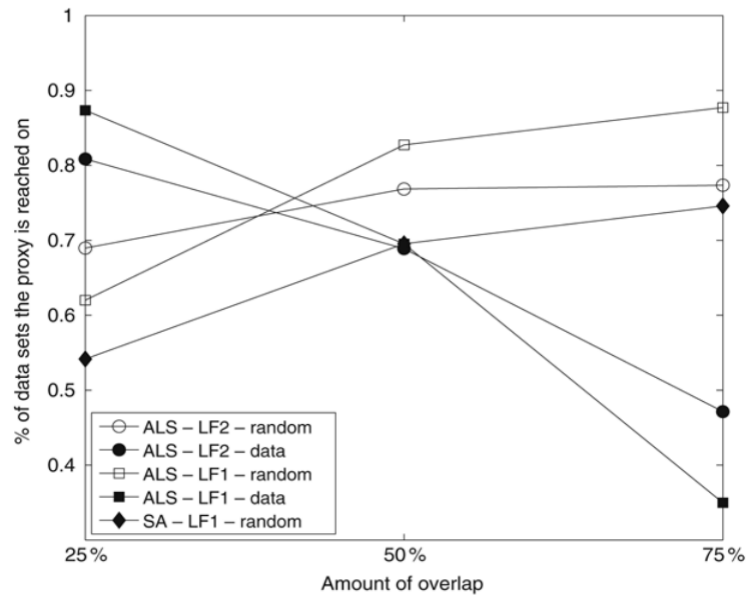


Figure 2.1: Comparison of additive overlapping algorithms (Depril et al., 2008)

2.3.4 Low dimensional additive overlapping clustering LAOC

Low additive overlapping clustering method is based on Mirkin's model which uses a set of vectors to describe the clusters' profiles :

$$M = AP \tag{2.13}$$

where M is the Merkin's model, A is the clusters memberships matrix and P represents the clusters profiles matrix. In the Mirkin's model, The interpretation of clusters is founded on the cluster profiles. It is as difficult as the number of the variables expands. Clustering can be gravely hampered if the data enclose a big number of variables (Mirkin, 1987). There are several methods that use the dimension reduction in order to deal with this problem. But these methods perform the dimension reduction and the clustering in a sequential alternating manner and cannot

perform both of them at the same time. LAOC (Depril et al., 2012) performs both of dimensional reduction and clustering simultaneously. The profiles of Merkin's model are computed using the following equation of P:

$$P = CB \tag{2.14}$$

$$S \leq min(K, J) \tag{2.15}$$

where S is the rank of P with P is the product of C the matrix of the profiles' scores and B the matrix containing the base vectors of low dimensional row space. The data X is approximated in the low additive overlapping clustering with the model matrix as follows:

$$X = M + E \tag{2.16}$$

where E represents the discrepancies between X and M.
The objective criterion function of the low additive overlapping clustering is defined as follows :

$$L^2(A, P) = \|X - AP\|_F^2 = \sum_{i,j=1}^{I,J} (x_{ij} - \sum_{k=1}^{K} a_{ik} p_{kj})^2 \tag{2.17}$$

Practically, the values of the number of clusters k and the number of the subspace S are unknown. For this reason, a series of models according to a range of values for S and k can be performed, one of the generated models can be chosen.

2.4 Geometrical overlapping clustering methods

Geometrical overlapping clustering methods consider the overlaps as a barycentre on the related clusters' representatives. Minimizing the sum of distances between the average of clusters representatives and each observation with the adoption of the geometrical reasoning. There are diverse geometrical overlapping clustering methods such as the overlapping k-means (Cleuziou, 2007), weighted overlapping k-means (Cleuziou, 2009), kernel overlapping k-means with its two variants KOKMI (Ben N'Cir et al., 2015), KOKMII (Ben N'Cir et al., 2015) and the two Parameterized overlapping k-means R_1OKM (Ben NCir et al., 2014) and R_2OKM (Ben NCir et al., 2014).

2.4.1 Overlapping K-means

Proposed by (Cleuziou, 2007) overlapping k-means aims to generate an appropriate classification into overlapping clusters. This method is based on the definition of an objective criterion that allows to estimate overlapping schemes, and an heuristic that allows to approach the best coverage considering the criterion. OKM is considered as a generalization of k-means method. The objective criterion is a generalization of the square error criterion used in the k-means. Analogously, the used heuristic is the generalization of the k-means algorithm. K-means algorithm is a linear time-complexity method that shows a good performance owing to its simplicity. In spite of the high complexity of the finding an optimal partition task. The k^n possible combinations to arrange n objects into k clusters defines the high computational requirements of the task. The k-means concept consists in the iteration of computation of the clusters centres and the assignment of each observation to the cluster with the nearest center, this procedure minimizes the square error criterion:

$$J_{K-means}(P) = \sum_{p_j \epsilon P} \sum_{x_i \epsilon P_j} d^2(x_i, p_j) \tag{2.18}$$

where p_j is the barycentre of the cluster j. The generated partition by k-means algorithm is a local optimum that depends on the initial partition and the choice of the number of clusters k. In the overlapping k-means algorithm, there exist $2^{k.n}$ possible combinations in which to arrange n observations into k overlapping clusters. Thus the overlapping k-means method proposes a new objective criterion in order to make it possible to detect overlapping clusters.

The objective criterion introduced in the OKM is an extension of the k-means square error criterion using two sets of parameters. A binary memberships matrix A to represent the assignments relationships of the objects with the clusters, and a matrix P containing the clusters centres. The objective criterion measures for each data, the sum of Euclidean distance that separates it from its representatives.

$$J_{OKM}(A, P) = \sum_{i=1}^{N} ||x_i - (\bar{x}_i)||^2 \tag{2.19}$$

$(\bar{x}_i)$ represents the barycentre of the clusters representatives. The global process of overlapping k-means is similar to k-means algorithm. The iteration of the assignment and the centres calculation is preceded by an initialization step. OKM algorithm introduces a new centres calculation way and new assignment procedure in OKM. The objective function's minimization is accomplished by alternating two processes: Calculation of clusters' representatives P, and the memberships assignments A. Multiple memberships assignment is realized using an heuristic exploring part of the combinatorial set of possible assignments. The heuristic arranges clusters

form the closest to the farthest for each observation. Then assigns the observations to the arranged clusters in order to minimize the distance between the observation and its representative. $(\bar{x}_i)$ represents the barycentre of clusters representatives where x_i is assigned and a_i is the set of clusters to which x_i is assigned, p_k is the representative of the cluster C.

$$(\bar{x_i}) = \sum_{k \in A_i} \frac{p_k}{|A_i|}. \tag{2.20}$$

Clusters representatives' update is done locally for each cluster. A_i defines the number set of observations assigned to the cluster K.

$$p_k^* = \frac{\sum\limits_{x_i \in a_k} \frac{1}{|a_i|^2} . \tilde{x_i^k}}{\sum\limits_{x_i \in a_k} \frac{1}{|a_i|^2}}. \tag{2.21}$$

OKM allows observations to belong to one or several clusters and aims to find a recovery A of K overlapping clusters relying on the Algorithm 2.4 which uses an assignment strategy that is described in the Algorithm 2.5.

Algorithm 2.4 Two-step reallocation process for overlapping clustering (Cleuziou, 2007)

1: Input: Data matrix X, number of clusters K, the objective function J , threshold on objective function ϵ, maximum number of iterations t_{max}.
2: initialize draw randomly k initial cluster prototypes ;
3: **FOR** each $x_i \in X$, derive the initial coverage A_0 from P_0;
4: Set $t = 0$;
5: **DO**
6: **FOR** each cluster, compute successively the new prototype P_k^{t+1};
7: **FOR** each $x_i \in X$ compute the new assignment A^{t+1} upon P_{t+1};
8: **WHILE** $(t < tmax)$or$(J_{OKM}^t - J_{OKM}^{t+1} > \epsilon)or(J_{OKM}^t > J_{OKM}^{t+1})$
9: Output: a binary membership matrix A, matrix of cluster profiles P.

2.4.2 Parametrized Overlapping K-means R_1OKM

Prior geometrical overlapping clustering methods generate overlapping clustering without the ability to supervise the amount of the overlap in the resulting structure. The overlap size in these methods is influenced by the model that the method adopts. The clustering method result must

Algorithm 2.5 Overlapping k-means assignment strategy (Cleuziou, 2007)

1: Input: X: data matrix, P: matrix of k cluster prototypes, J: OKM objective function, $\tilde{A}_i$: the previous assignment for x_i.
2: Look for the index q of the nearest cluster profile of x_i :
$q = argmin_{k\in[1;K]} \|x_i - A_i P_k\|$;
3: Set $A_{i,k} = 0, \forall k \neq q$;
4: Set $A_{i,q} = 1$;
5: **DO**
6: Look for the index q of the nearest cluster profile of x_i :
$q = argmin_{k\in[1;K]\wedge A_{i,k}=0} \|x_i - A_i P_k\|$;
$A_i^{'} = A_i$ and $A_i^q = 1$
7: **IF** $(J_{OKM}(A_i^{'}, P) \leq J_{OKM}(A_i, P))$ **THEN** $A_i = A_i^{'}$;
8: **ELSE** $q = 0$;
9: **WHILE** $(q > 0 and \|A_i]\| < K)$
10: **IF** $(J_{OKM}(A_i, P) \leq J_{OKM}(\tilde{A}_i, P))$ **THEN** return A_i
11: Output: A_i : a binary membership vector .

satisfy the convenience to the data. Therefore, (Ben NCir et al., 2014) proposed the parametrized overlapping k-means (R_1OKM). A geometrical overlapping clustering method that allows overlap regulation based on the overlapping k-means (Cleuziou, 2008) taking into account the number of assignment of the implicated cluster. The main concept of the overlapping k-means is to assign a data observation x_i to the closest set of cluster profiles while minimizing its objective criterion. To decide whether to select a first cluster assignment combination a_i for the observationx_i or a second cluster assignment combination $a_i^{'}$ represented by binary vector. A positive difference in the produced local error is claimed. The main concept of the Parametrized OKM is to control the cluster assignment choice decision with the insertion of weights on the local errors relying on the size of the combinations. This weighting is manipulated using a parameter α. The assignment decision is formalized as follows:

$$\left\|A_i^{'}\right\|^{\alpha} . \left\|x_i - p^{A_i}\right\|^2 - \|A_i\|^{\alpha} . \left\|x_i - p^{A_i^{'}}\right\|^2 > 0 \tag{2.22}$$

where p^{A_i} represents the combination $p^{A_i} = \sum_k a_{i,k} p_{k,j} / \|A_i\|$.
The objective criterion of R_1OKM based on the Overlapping K-means objective criterion:

$$J_{R_1OKM}(A, P) = \sum_i \left[\sum_k a_{i,k}\right]^{\alpha} \sum_i \left(x_{i,j} - \frac{\sum_k a_{i,k} p_{k,j}}{\sum_k a_{i,k}}\right) \tag{2.23}$$

where K is the predetermined number of clusters. The model behaves in function of the value of α. If $\alpha = 0$, it extinguishes the effect of the weighting on the combination sizes. Therefore, when $\alpha = 0$, the obtained result is the same as the result of the OKM. When $\alpha > 0$, it restricts the assignment with wide combinations. As high as α rises, the overlaps vanishes until the apparition of an identical clustering to the classic k-means. But when $\alpha < 0$, it supports the overlaps of the large assignments. As long as α bends down, the overlaps increases until the arise of a clustering where all data objects are assigned to every cluster. The obtained cluster profile is a weighted mean of the cluster profiles according to each data belonging to the cluster. The Parameterized ROKM method follows the same reallocation process of the overlapping k-means method with the use of the objective criterion of R_1OKM instead of OKM's objective criterion. The cluster profiles update is computed for each cluster in a successive manner depending on the membership matrix and the residual values of the profiles matrix P in order to extract the optimal cluster j profile $p^*_{k,j}$:

$$p^*_{k,j} = \frac{\sum_i a_{i,k}(\sum_l a_{i,l})^{\alpha-2} p^i_{k,j}}{\sum_i a_{i,k}(\sum_l a_{i,l})^{\alpha-2}} \tag{2.24}$$

where P^i_k defines the perfect cluster profile P_k in accordance with the data object x_i. The local error on x_i is zero:

$$p^i_{k,j} = x^{i,j} \sum_l a_{i,l} - \sum_{l \neq k} a_{i,l} p_{l,j} \tag{2.25}$$

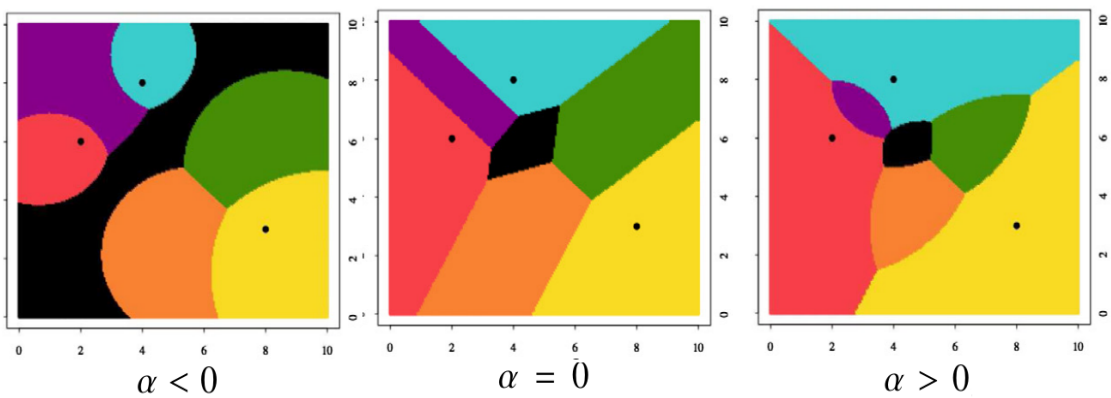


Figure 2.2: Example of voronoi cells induced by the R1OKM model (Ben NCir et al., 2014)

The Parametrized OKM method offers the possibility to regulate the overlap size according to the number of assignments of the cluster in order to produce a closer clustering to the expected

shape of the data by the expert.
Figure 2.2, (Ben NCir et al., 2014) shows the effect of the overlap regulation using the parameterized OKM with $\alpha = 0$, $\alpha > 0$ and $\alpha < 0$

2.4.3 Parametrized overlapping K-means R_2OKM

Proposed by (Ben NCir et al., 2014), after R_1OKM that regulates overlaps considering the number of assignments. R_2OKM is another parametrized overlapping k-means method that represents a model based on the dispersal of the cluster profiles with the observation to be assigned. Aiming to allow control on the overlap quality, it assumes that the regulation of the overlap have to perform in different manners regarding the area of the cluster. While the main idea of the R_2OKM considers that when overlaps are required to be reduced, the major anticipation is to disconnect more the distant clusters in order to isolate them. The following function formalizes the concept of the overlap regulation of the R_2OKM in order to promote or to restrict the local error between each data object and the cluster involved in the assignment alternatives. This function represent the dispersal criterion measuring the square distance from the data object x_i to the clusters profiles where it belongs. This measure can be boosted or limited with the use of a parameter λ in the objective criterion, the dispersal criterion is defined as follows:

$$D_{A,P}(x_i) = \frac{\sum_k a_{ik} \sum_j (x_{ij} - p_{kj})^2}{\sum_k a_{ik}} \tag{2.26}$$

The new objective criterion proposed by (Ben NCir et al., 2014)) manipulates the overlap by introducing a parameter λ which affects the dispersal criterion between the data and the clusters to help to obtain the expected overlap, the objective criterion of the R_2OKM is defined as follows:

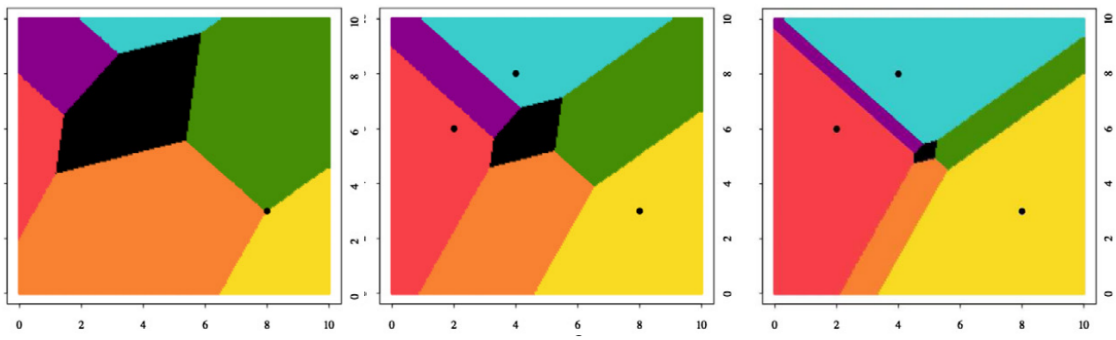
$$J_{R_2OKM}(A, P) = \sum_i \sum_j \left[\left(x_{ij} - \frac{\sum_k a_{ik} p_{kj}}{\sum_k a_{ik}} \right) + \lambda . D_{A,P} \right] \tag{2.27}$$

where λ is the parameter that helps to handle the overlaps. Affecting the dispersal criterion, when $\lambda = 0$ the overlap regulation have no more effect. The obtained clustering is identical to the OKM clustering result. But when $\lambda > 0$, it restricts the overlap. The overlap decreases as long as λ increases until it vanishes . When $\lambda < 0$, it stimulates the size of the overlap. The overlap expands as long as λ decreases until the emergence of a clustering where every data object is assigned to all the clusters. The objective criterion helps to direct the reassignment process in the usual OKM algorithm. The algorithm of the R_2OKM is similar to the OKM algorithm 2.4 but it

uses the R_2OKM objective criterion in the assignment heuristic. Clusters profiles are updated in a sequential manner for each cluster as defined in the following function :

$$P^*_{kj} = \frac{\sum_i a_{i,k} \left[\left(\sum_l a_{i,l}\right)^{-2} p^i_{kj} + \lambda \left(\sum_l a_{i,l}\right)^{-1} x_{i,j} \right]}{\left[\sum_i a_{i,k} \left[\left(\sum_l a_{i,l}\right)^{-2} + \lambda \left(\sum_l a_{i,l}\right)^{-1} \right] \right]} \tag{2.28}$$

R_2OKM is considered as a generalization of the C-means algorithm and the OKM algorithm. As for the OKM algorithm, the used technique for the profiles updating and for the assignments garanties the convergence and are characterized with a low cost in complexity in $O(TNKLogK)$ where T is the number of iterations in the reallocation process. The minimization process of the objective criterion function in the different steps may induce a local optimum as result. Therefore, this issue can be solved in practice by estimating several clusters analysis using different profiles initializations. The clustering showing the lowest value of the optimized criterion function will be preserved



$\lambda < 0$ $\lambda = 0$ $\lambda > 0$

Figure 2.3: Example of voronoi cells induced by the R2OKM model (Ben NCir et al., 2014)

Figure 2.3 shows the effect of the overlap regulation using the R_2OKM with $\lambda = 0$, $\lambda > 0$ and $\lambda < 0$

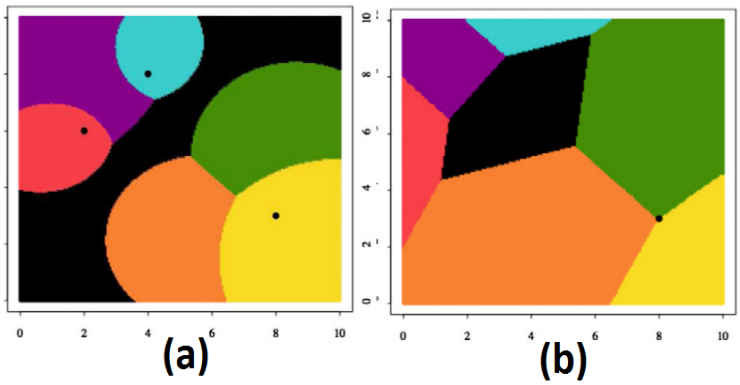


Figure 2.4: Example of voronoi cells induced by the R1OKM and the R2OKM models (Ben NCir et al., 2014)

Figure 2.4 shows how the shapes of the obtained clusters with the R1OKM (a) are different from those obtained with R2OKM (b) for $\alpha < 0$ and $\lambda < 0$.

2.5 Conclusion

In this chapter, we presented an overview of hard partitional overlapping clustering methods such as additive and geometrical methods. Additive methods consider that the addition of the representatives of the concerned clusters is the overlap results while geometrical methods consider the overlaps as a barycentre on the overlapping clusters. However, these methods present an issue when a data structure is expected with predefined size of overlap. In the next chapter, we will present the two additive methods that we propose to introduce identifying overlapping clusters with control of overlap's size.

Chapter 3

Overlap regulation for additive overlapping clustering methods

3.1 Introduction

Overlapping clustering is a knowledge discovery process that allows the detection of overlapping clusters from an unlabelled dataset. However, overlap regulation allows to obtain overlapping clusters in accordance with the expected overlap size. The concept of overlap regulation have been introduced only with geometrical methods such as Parametrized ROKM and R2OKM. Therefore, in this chapter we aim to introduce the overlap regulation into two additive methods ALS_{lf2} and ALS_{lf1}.

This chapter is structured as follows: In Section 3.2, we present the motivation behind the development of our proposed solution. In Section 3.3, we present the alternating least square type (ALS). Section 3.4 details the proposed methods RALS1 and RALS2. Finally, we describe a step by step example of the proposed methods in section 3.5.

3.2 Motivation

Overlapping clustering allows data objects to belong to several groups in order to obtain an overlapping clustering. Most of the existing overlapping clustering methods produce clusters with static and large overlaps sizes. The fact that overlapping clustering methods produce more information than needed, reduces the precision of the resulting classification. Which affects the quality of the obtained clustering. The more the method produces larger clustering than needed, the more the obtained clustering is inefficient. The size of the overlap depends on the data structure, the used clustering method and the number of clusters and several other factors. However, the expert should be able to control the size of the overlap when an overlapping structure is expected. Clusters with a larger overlap than needed are not appropriate for clustering applications. The overlap size should be regulated according to the clustering application's requirements. For that reason, we study overlap sizes generated by the existing ALS_{lf1} and ALS_{lf2} methods. Figures 3.2 and 3.3 show an example of a clustering obtained by ALS_{lf1} and ALS_{lf2} methods. We notice the large overlapping part which is not appropriate to the expected data structure in Figure 3.1.

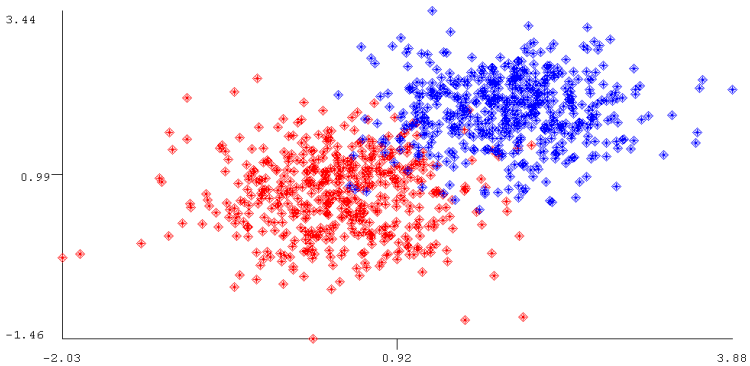


Figure 3.1: Expected clustering structure

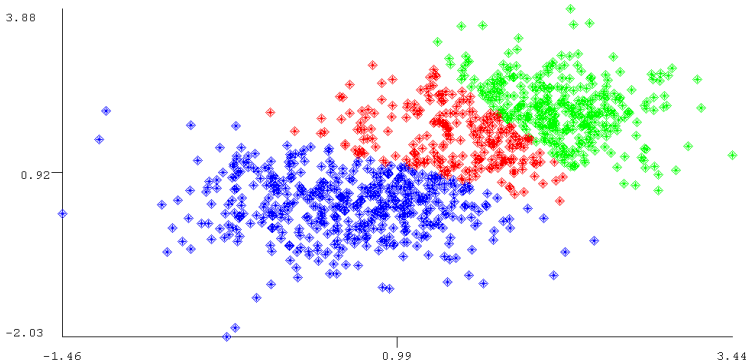


Figure 3.2: Obtained clustering with ALS_{lf1}

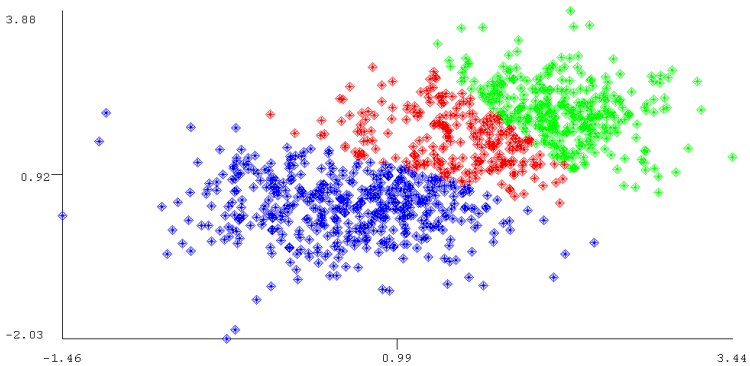


Figure 3.3: Obtained clustering with ALS_{lf2}

The expected clusterings are characterized with different rates of overlaps according to the characteristics of the concerned application. The regulation control is proposed only with geometrical methods offering the manipulation of the size of overlap in order to obtain the anticipated overlapping clustering result. Up to now, there are no additive overlapping clustering methods that introduce the overlap regulation principle. Therefore, we focus our study on ALS_{lf1} and ALS_{lf2} and on the additive model which will be generalized in the purpose to produce overlap regulated clusterings.

Therefore, we aim to introduce the possibility to regulate the overlap size with both of the ALS_{lf2} and ALS_{lf1} methods.

3.3 Alternating least square (ALS)

Alternating least square algorithms types are based on the one mode additive clustering model (MIRKIN, 1990). The one-mode additive clustering approximates the matrix X by model matrix M with $I \times J$ entries that can be decomposed into an $I \times K$ membership matrix A and a $K \times J$ profiles matrix P.

$$m_{ij} = \sum_{k=1}^{K} a_{ik} p_{kj} \tag{3.1}$$

where m_{ij} represents the ith row element of M, the columns of the membership matrix A represent K clusters, the value of a_{ik} indicate if the object i belongs to the cluster or not. When $a_{ik} = 0$ the object i is not assigned to the cluster k and $a_{ik} = 1$ means that the object i belongs to the cluster k. Zero columns are not allowed in A which is the only restriction put on the membership matrix. This allows the resulting clustering to be an overlapping clustering. The rows of the profiles matrix P represents the features of profiles for each cluster p_{kj} indicates the jth feature value of the cluster k. Those values represent the sum of the profiles of the assigned objects to the cluster k. There exist some discrepancies in between the model matrix M and the data matrix X which means that:

$$X = M + E \tag{3.2}$$

where $M = AP$. By replacing M by AP in Equation 3.2 leads to $X = AP + E$. Which means that $E = X - AP$ the discrepancies between X and M. Therefore the model matrix M is estimated by minimizing the loss function of least square over A and P

$$lf(A, P) = \|X - AP\|_F^2 = \sum_i \sum_j \left(x_{ij} - \sum_k a_{ik} p_{kj} \right)^2 \tag{3.3}$$

where $\|.\|_F^2$ indicates the Frobenius norm of a matrix which is the square root of the squared entries of the matrix. When the membership matrix is given, the optimal cluster profiles matrix P represents the least squares multiple regression estimator:

$$P = (A'A)^{-1} A'X \tag{3.4}$$

By replacing Equation 3.4 in Equation 3.3 an alternative of the loss function appears with only one parameter:

$$lf(A) = \left\| X - A(A'A)^{-1} A'X \right\|_F^2 \tag{3.5}$$

Estimating Equation 3.5 consists of the exploration of all the 2^{IK} possible binary membership combinations.

3.4 Additive overlapping clustering methods with overlap regulation

In the purpose to make overlap regulation for additive model, we propose the $RALS\,1$ and the $RALS\,2$ methods. Which formalizes an overlap regulation principle based on the number of assignment of the data objects. The proposed regulation process aims to parametrize the expected benefit of an overlap on the local errors. The large overlaps generated by ALS_{lf1} and ALS_{lf2} are induced by the model that favours the assignment of a data x_i to the nearest combination of clusters profiles as described in Equation 2.3.

The proposed solution include two new methods $RALS\,2$ and $RALS\,1$ by introducing the overlap principle respectively in ALS_{lf2} and ALS_{lf1}.

3.4.1 Proposed RALS1 method

In RALS1 methods, The decision on choosing the assignment combinations is performed depending on the induced local error between two assignment combinations A_i and $A_i^{'}$. The proposed overlap regulation principle consists in weighting local error and adjusting the weight of the combinations with a parameter α.
The assignment decision is regulated by the following expression :

$$\left\|A_i^{'}\right\|^{\alpha} . \left\|x_i - p^{A_i^{'}}\right\|^2 - \|A_i\|^{\alpha} . \left\|x_i - p^{A_i}\right\|^2 > 0. \tag{3.6}$$

where p^{A_i} determines the combination $p^{A_i} = \sum_k a_{ik} p_{kj}$.
The new model favours or limits the assignment of x_i to the clusters depending on the value of α. The overlap regulation model results in the following objective criterion:

$$J_{RALS\,1}(A, P) = \sum_i \left[\sum_k a_{ik}\right]^{\alpha} \sum_j \left(x_{ij} - \sum_k a_{ik} p_{kj}\right)^2 . \tag{3.7}$$

The behaviour of the model depends on the value of α : When $\alpha = 0$, there is no overlap regulation. Therefore the clustering result is equivalent the clustering obtained with the original $ALS_l f1$ method. When $\alpha > 0$, it restricts the large combinations assignments. As high as α increases, the overlap size decreases until there is no more overlap in the clustering result. But, when $\alpha < 0$ it favours the assignment of large combinations. The overlap size increases as long

as α decreases until the apparition of a clustering result where each object belongs to any cluster.

The algorithm begins with an initial profile matrix P_0 that seeds the algorithm can be randomly obtained or given by the user. Once the initial cluster profiles matrix obtained, the algorithm computes the initial membership matrix A_0 conditionally optimal upon P_0 according to the assignment strategy. In the assignment strategy 3.2, the algorithm explores for each data object every possible combination of clusters and assigns the combination that produces the lowest loss value. The freshly assigned combination is inserted immediately in the membership matrix A as shown in . After each row assignment, The *RALS* 1 algorithm computes immediately the profiles matrix P conditionally optimal upon A. Once all the rows of A are estimated, the algorithm compares the loss function obtained by the new membership matrix with the old one and picks up the one that yields the less loss function.
RALS 1 Algorithm produces 3.1 a set of membership matrices where each one induce a lower value of loss function than the preceding one. Thus, we obtain a set of decreasing loss values. The obtained set converges due to the finite number of membership matrices. The profiles matrix update Upon A is computed as for the original ALS_{lf1} profiles computation:

$$P = (A^{'}A)^{-1}A^{'}X. \tag{3.8}$$

There is no weighting on the clusters profiles matrix. When there exist a zero column in the membership matrix A, the inverse of AA does not exist. Therefore we use the Moore-Penrose pseudo-inverse in order to calculate the least squares profiles. The quality of clustering result produced by the RALS1 algorithm depends hardly on the initial matrices A_0 or P_0. In order to avoid the local optimum trap multi-start procedure is used and the solution generating the lowest loss value is retained.

Algorithm 3.1 RALS1 pseudo-code

Require: X: Data matrix.
K : number of clusters.
it_{max} : number of maximum iterations.
ϵ : minimal improvement in the objective function.

1: Initialize cluster profiles P_0 randomly.
Initialize cluster memberships A_0 over P_0.
Compute objective function (A_0, P_0).
2: it = it +1.
3: Compute new assignment A_{it} using the AS1(X,P_{it},A_{it-1}).
4: Compute objective function $J(A_{it}, P_{it})$.
5: **if** $it \leq it_{max}$ and $(J(A_{it-1}, P_{it-1}) - J(A_{it}, P_{it}) > \epsilon)$. **then**
6: Go to 2.
7: **else**
8: Return A_{it} the final membership matrix.
9: **end if**

Algorithm 3.2 Assignment Strategy AS1

Require: X: Data matrix.
K : number of clusters.
P : Profiles matrix.
A : membership matrix.

1: **for all** $x_i \in X$ **do**
2: Generate all possible 2^K cluster assignment combination
3: **for all** combinations **do**
4: Compute induced loss error using Equation 3.9
5: **end for**
6: Assign the combination inducing the minimal loss value into the ith row of A
7: Update clusters profiles P_{it} using Equation 3.4.
8: **end for**
9: Return A the Estimated membership matrix.

RALS 1 estimates the cluster profiles after each row assignment in A. Which characterizes *RALS* 1 with a high complexity. For that reason, in the next section we propose to introduce the

overlap regulation principle by $RALS2$ which generalizes ALS_{lf2} which is characterized with lower complexity level than $RALS1$.

3.4.2 Proposed RALS2 method

A second new algorithm, called $RALS2$, is also based on the same overlap regulation principle defined in Equation 3.6. $RALS2$ algorithm minimizes the loss function 3.7 by assigning for each row of A the assignment combination that minimizes the most the local error. The profiles matrix that seeds RALS2 algorithm can be a priori given by the user or randomly obtained. Once the initial profiles matrix P_0 obtained, the algorithm estimates A_0 conditionally optimal upon P_0. Each row is estimated by enumerating all the possible combinations. The combination that generates the lowest local error. The parameter α regulates the weights of multiple assignment combinations. $\alpha = 0$ cancels the overlap regulation, $\alpha > 0$ restricts overlapping assignment combinations while $\alpha < 0$ favours overlapping assignment combinations. Assignment strategy 3.4 assigns the combination that induces the minimal loss error:

$$lf(A,P) = \sum_j \left(x_{1j} - \sum_{k=1}^{K} a_{1k} p_{kj} \right)^2 + \cdots + \sum_j \left(x_{Ij} - \sum_{k=1}^{K} a_{Ik} p_{kj} \right)^2 \tag{3.9}$$

This signifies that the intervention of the ith row of A_1 to the loss function can be segregated from the roles played by the other rows. When all the rows are estimated, the algorithm estimates the new profiles matrix and compares the loss function obtained the new assignment matrix with the old one. Then, the algorithm verifies if a decrease have been established. In which case, the row wise updating procedure is repeated on the new membership matrix after computing the new conditionally optimal profiles matrix upon the recently obtained membership matrix. The algorithm alternates between the computation of membership matrices A upon profiles matrices P and P upon A conditionally optimal to each others. The alternation cycle stays in action until one of the following condition is fulfilled:

- The iteration number reaches the maximum number of iterations specified by the user.
- The difference between the obtained loss function and the previous one is lower than the minimal improvement constraint given by the user.
- There is no further decrease in the loss function after the assignment step.

Anageously to *RALS* 1, the clusters' profiles matrix is updated using Equation 3.8. The main difference between*RALS* 1 and *RALS* 2 is the profiles matrix upload as shown in algorithms 3.1 and 3.3. In *RALS* 2 the profiles matrix is uploaded after the estimation of all the rows of the membership matrix. While in *RALS* 1 the profiles matrix is uploaded immediately after each row assignment in the membership matrix. Which allows to reduce the high complexity level that characterizes *RALS* 1. The obtained result by *RALS* 2 algorithm depends on A_0. Therefore, a multi-start with different initializations is recommended in order to avoid local optimum solutions.

Algorithm 3.3 RALS2 pseudo-code

Require: X: Data matrix.
K : number of clusters.
it_{max} : number of maximum iterations.
ϵ : minimal improvement in the objective function.
1: Initialize cluster profiles P_0 randomly.
Initialize cluster memberships A_0 over P_0.
Compute objective function (A_0, P_0).
2: it = it +1.
3: Update clusters profiles P_{it} using Equation 3.4.
4: Compute new assignment A_{it} using the AS2(X,P_{it},A_{it-1}).
5: Compute objective function $J(A_{it}, P_{it})$.
6: **if** $it \leq it_{max}$ and $J(A_{it-1}, P_{it-1}) - J(A_{it}, P_{it}) > \epsilon$. **then**
7: Go to 2.
8: **else**
9: Return A_{it} the final membership matrix.
10: **end if**

Algorithm 3.4 Assignment Strategy AS2

Require: X: Data matrix.
K : number of clusters.
P : Profiles matrix.
A : membership matrix.

```
for all x_i ∈ X do
    generate all possible 2^K cluster assignment combination
    for all combinations do
        Compute induced loss error using Equation 3.9
    end for
    assign the combination inducing the minimal loss value into the i^th row of A
end for
Return A_it the final membership matrix.
```

3.5 Step by step example for *RALS* 2

Given the following dataset described in Table 3.1, we use *RALS* 2 to look for two groups. First we run RALS2 with $\alpha = -1.5$, Then we run RALS2 with $\alpha = 1.5$.

Table 3.1: Example Dataset X

Objects	a_1	a_2
x_1	2	4
x_2	3	2
x_3	2	1
x_4	1	6
x_5	4	5
x_6	1	4
x_7	6	6

3.5.1 Execution of $RALS2$ with $\alpha = -1.5$

Initialization of parameters: The maximum number of iterations is set to three iterations $t_{max} = 3$. The minimum improvement criterion is fixed to $\epsilon = 0,1$. the number of clusters is $K = 2$. The overlap regulation parameter is fixed to $= -1.5$.
Cluster profiles are initialized as shown in table 3.2.

Table 3.2: Clusters' profiles

Clusters	a_1	a_2
p_1	1	5
p_2	5	1

Estimation of the memberships matrix: The estimation of the membership matrix is performed row by row for each observation by enumerating all the possible combinations. In this example, there are two clusters. Three possible assignment combinations are available for each row.
For the observation x_1:
The relative loss for each combination according to $l(combination) = [a_{1k}]^{\alpha} \sum (x_{1j} - p_{kj})^2$

- for the combination 1: $l(1) = 1^{-1.5}.\left((2-1)^2 + (4-5)^2\right) = 2$
- for the combination 2: $l(2) = 1^{-1.5}.\left((2-5)^2 + (4-1)^2\right) = 18$
- for the combination 3: $l(3) = 2^{-1.5}.\left((2-(5+1))^2 + (4-(1+5))^2\right) = 7$

The combination generating the small value of loss is chosen and the object is assigned according to the chosen combination 0,1. The observation x_1 is assigned to the cluster 1.

For the observation x_2:

- for the combination 1: $l(1) = 1^{-1.5}.\left((3-1)^2 + (2-5)^2\right) = 13$
- for the combination 2: $l(2) = 1^{-1.5}.\left((3-5)^2 + (2-1)^2\right) = 5$

- for the combination 3: $l(3) = 2^{-1.5}.\left((3-(5+1))^2 + (2-(1+5))^2\right) = 8.75$

The observation x_2 is assigned to the cluster 2.

For the observation x_3:

- for the combination 1: $l(1) = 1^{-1.5}.\left((2-1)^2 + (1-5)^2\right) = 17$
- for the combination 2: $l(2) = 1^{-1.5}.\left((2-5)^2 + (1-1)^2\right) = 9$
- for the combination 3: $l(3) = 2^{-1.5}.\left((2-(5+1))^2 + (1-(1+5))^2\right) = 14$

The observation x_3 is assigned to the cluster 2.

For the observation x_4:

- for the combination 1: $l(1) = 1^{-1.5}.\left((1-1)^2 + (6-5)^2\right) = 1$
- for the combination 2: $l(2) = 1^{-1.5}.\left((1-5)^2 + (6-1)^2\right) = 51$
- for the combination 3: $l(3) = 2^{-1.5}.\left((1-(5+1))^2 + (6-(1+5))^2\right) = 8.75$

The observation x_4 is assigned to the cluster 1.

For the observation x_5:

- for the combination 1: $l(1) = 1^{-1.5}.\left((4-1)^2 + (5-5)^2\right) = 9$
- for the combination 2: $l(2) = 1^{-1.5}.\left((4-5)^2 + (5-1)^2\right) = 17$
- for the combination 3: $l(3) = 2^{-1.5}.\left((4-(5+1))^2 + (4-(1+5))^2\right) = 2.8$

The observation x_5 is assigned to both of the clusters 1 and 2.

For the observation x_6:

- for the combination 1: $l(1) = 1^{-1.5}.\left((1-1)^2 + (4-5)^2\right) = 1$
- for the combination 2: $l(2) = 1^{-1.5}.\left((1-5)^2 + (4-1)^2\right) = 17$
- for the combination 3: $l(3) = 2^{-1.5}.\left((1-(5+1))^2 + (4-(1+5))^2\right) = 10.15$

The observation x_6 is assigned to the clusters 1.

For the observation x_7:

- for the combination 1: $l(1) = 1^{-1.5}.\left((6-1)^2 + (6-5)^2\right) = 26$
- for the combination 2: $l(2) = 1^{-1.5}.\left((6-5)^2 + (6-1)^2\right) = 26$
- for the combination 3: $l(3) = 2^{-1.5}.\left((6-(5+1))^2 + (6-(1+5))^2\right) = 0$

The observation x_5 is assigned to both of the clusters 1 and 2.

Table 3.3: Assignment matrix

A/P	p_1	p_2
x_1	1	0
x_2	0	1
x_3	0	1
x_4	1	0
x_5	1	1
x_6	1	0
x_7	1	1

Table 3.3 shows how each data object is assigned to a combination of clusters when the parameter $\alpha = -1.5$. The loss function is the sum of the losses obtained by each row estimation.

$$J_{RALS2}(A, P) = \sum_i \left[\sum_k a_{ik}\right]^{\alpha} \sum_j \left(x_{ij} - \sum_k a_{ik} p_{kj}\right)^2 = 20.8. \tag{3.10}$$

Once the assignment matrix estimated, the profiles matrix are computed upon A using $P = (A'A)^{-1}A'X$. after the profiles matrix estimation, the algorithm re-estimates the membership matrix upon the new profiles matrix and compares the obtained loss value with the old one. The final clustering is obtained when one of the conditions is fulfilled i.e. the iteration number reaches t_{max}, or when there is no further decrease in the loss function.

3.5.2 Execution of RALS2 with $\alpha = 1.5$

Initialization of parameters: The same parameters initialization is adopted in this example except of the overlap parameter $= 1.5$.

Estimation of the memberships matrix: The estimation of the membership matrix is performed row by row for each observation by enumerating all the possible combinations.

For the observation x_1:
The relative loss for each combination according to $l(combination) = [a_{1k}]^{\alpha} \sum (x_{1j} - p_{kj})^2$

- for the combination 1: $l(1) = 1^{1.5}.\left((2-1)^2 + (4-5)^2\right) = 2$
- for the combination 2: $l(2) = 1^{1.5}.\left((2-5)^2 + (4-1)^2\right) = 18$
- for the combination 3: $l(3) = 2^{1.5}.\left((2-(5+1))^2 + (4-(1+5))^2\right) = 56.4$

The combination generating the lowest value of loss is chosen and the object is assigned according to the chosen combination 0,1. The observation x_1 is assigned to the cluster 1.

For the observation x_2:

- for the combination 1: $l(1) = 1^{1.5}.\left((3-1)^2 + (2-5)^2\right) = 13$
- for the combination 2: $l(2) = 1^{1.5}.\left((3-5)^2 + (2-1)^2\right) = 5$
- for the combination 3: $l(3) = 2^{1.5}.\left((3-(5+1))^2 + (2-(1+5))^2\right) = 70.5$

The observation x_2 is assigned to the cluster 2.

For the observation x_3:

- for the combination 1: $l(1) = 1^{1.5}.\left((2-1)^2 + (1-5)^2\right) = 17$
- for the combination 2: $l(2) = 1^{1.5}.\left((2-5)^2 + (1-1)^2\right) = 9$
- for the combination 3: $l(3) = 2^{1.5}.\left((2-(5+1))^2 + (1-(1+5))^2\right) = 115.62$

The observation x_3 is assigned to the cluster 2.

For the observation x_4:

- for the combination 1: $l(1) = 1^{1.5}.\left((1-1)^2 + (6-5)^2\right) = 1$
- for the combination 2: $l(2) = 1^{1.5}.\left((1-5)^2 + (6-1)^2\right) = 51$
- for the combination 3: $l(3) = 2^{1.5}.\left((1-(5+1))^2 + (6-(1+5))^2\right) = 70.5$

The observation x_4 is assigned to the cluster 1.

For the observation x_5:

- for the combination 1: $l(1) = 1^{1.5}.\left((4-1)^2 + (5-5)^2\right) = 9$
- for the combination 2: $l(2) = 1^{1.5}.\left((4-5)^2 + (5-1)^2\right) = 17$

- for the combination 3: $l(3) = 2^{1.5}.\left((4-(5+1))^2 + (4-(1+5))^2\right) = 22.56$

The observation x_5 is assigned to the clusters 1

For the observation x_6:

- for the combination 1: $l(1) = 1^{1.5}.\left((1-1)^2 + (4-5)^2\right) = 1$
- for the combination 2: $l(2) = 1^{1.5}.\left((1-5)^2 + (4-1)^2\right) = 17$
- for the combination 3: $l(3) = 2^{1.5}.\left((1-(5+1))^2 + (4-(1+5))^2\right) = 81.78$

The observation x_6 is assigned to the clusters 1.

For the observation x_7:

- for the combination 1: $l(1) = 1^{1.5}.\left((6-1)^2 + (6-5)^2\right) = 26$
- for the combination 2: $l(2) = 1^{1.5}.\left((6-5)^2 + (6-1)^2\right) = 26$
- for the combination 3: $l(3) = 2^{1.5}.\left((6-(5+1))^2 + (6-(1+5))^2\right) = 0$

The observation x_5 is assigned to both of the clusters 1 and 2.

Table 3.4: Assignment matrix

A/P	p_1	p_2
x_1	1	0
x_2	0	1
x_3	0	1
x_4	1	0
x_5	1	0
x_6	1	0
x_7	1	1

Table 3.4 shows how each data object is assigned to a combination of clusters when the parameter $\alpha = 1.5$. The loss function is the sum of the losses obtained by each row estimation.

$$J_{RALS2}(A, P) = \sum_i \left[\sum_k a_{ik} \right]^{\alpha} \sum_j \left(x_{ij} - \sum_k a_{ik} p_{kj} \right)^2 = 27. \tag{3.11}$$

This example shows clearly the impact of the value of parameter α on the multi-assignment decisions.

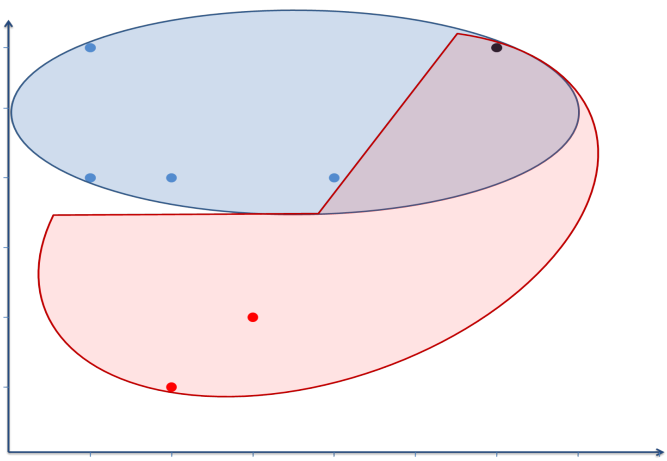


Figure 3.4: *RALS*2 process with $\alpha = 1.5$

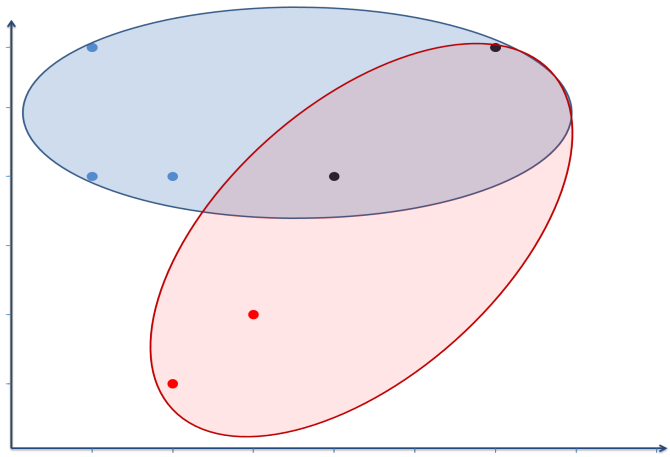


Figure 3.5: *RALS* 2 process with $\alpha = -1.5$

Figures 3.5 and 3.4 shows how the data objects are clustered in both of the two cases ($\alpha = 1.5$ and $\alpha = -1.5$). The blue data objects represent the first cluster and the red data objects represent the second cluster while the black data objects represent the objects belonging to the first and the second cluster. Figures 3.5 and 3.4 clearly shows how additive methods perspective consider overlapping data objects as the sum of the cluster profiles in question.

3.6 Conclusion

In this chapter,we described the alternating least square type on which are based the two proposed methods *RALS* 1and *RALS* 2. We described the algorithms of *RALS* 1 and *RALS* 2 which allow to introduce the overlap regulation based on the number of assignment in additive partitional overlapping clustering methods. Moreover, we presented the manner in which the overlap is manipulated. A step by step example of the proposed methods has been presented in this chapter in order to detail the overlap regulation principle.
The next chapter presents the experimental study that we have performed in order to validate the performance of our proposed methods and describes the evaluation measures that we have used on this purpose.

Chapter 4

Experimental study

4.1 Introduction

In this chapter, we present the evaluation measures that takes into consideration the assignment on multiple clusters. We also present the test datasets on which the experiments were preformed. We detail the different experiments which we have performed in order to evaluate the effectiveness of the overlap regulation with the proposed two methods *RALS* 1 and *RALS* 2 compared with existing geometrical and additive methods.
This chapter is organized as follows : Section 4.2 represents the adopted evaluation methodology that fits with overlapping clustering. Section 4.3 presents an evaluation of the results' quality relative of the two proposed methods on real and artificial datasets. Section 4.4 presents the results discussion.

4.2 Evaluation methodology

In the literature, there is a large set of evaluation methods that allows to compare the quality of a clustering algorithm. These methods are divided into two dominant families which are intrinsic metrics and extrinsic metrics. Intrinsic metrics are the metrics based on the closeness of elements belonging to one cluster to each others. Extrinsic metrics are the metrics based on comparison between the output of a clustering algorithm and classification built by human

assessors. In this work we will focus on the extrinsic metrics which are the most commonly used in clustering evaluation. Several evaluation metrics have been proposed, such as Purity and inverse Purity, Jaccard coefficient and mutual information. (Amigó et al., 2009) proposed formal constraints to compare extrinsic metrics where clustering evaluation should satisfy basic formal constraints such as cluster homogeneity and cluster completeness. In his study, (Amigó et al., 2009) compared the performance of several known metrics families. The study concluded that only BCubed metrics satisfy all of the formal constraints.
In clustering, each object is assigned to one cluster, while in overlapping clustering, each object can be assigned to several clusters. Thus, there is a problem with the evaluation of overlapping clustering algorithms with usual evaluation metrics. Standard evaluation metrics does not take into account the multiplicity of pairs of observations. An extended BCubed metric was proposed (Amigó et al., 2009) in order to deal with the problem of overlapping clustering evaluation. In the following we detail the extended BCubed metrics which we will use to evaluate the performance of our proposed solutions.

4.2.1 BCubed measures

The standard BCubed metrics (Bagga & Baldwin, 1998) decomposes the process of evaluation estimating the recall and the precision relative to each object in the distribution. BCubed metrics represents the effectiveness of the clustering system. The BCubed recall relative to an observation shows how many objects of its class are assigned to the same cluster. The BCubed precision relative to an observation shows how many objects of its cluster belong to the same class. Once an observation is accessed, the rest of data observations in the cluster can be explored. If the object is characterized with a high BCubed recall, most of related observations in the cluster are explored. If an object is characterized with a high BCubed precision, the noisy objects in the cluster are minimal. In BCubed metrics, $C_{(o)}$ and $P_{(o)}$ represent respectively the class and the cluster of an object. The correctness of a relation between o and $o^{'}$ in the distribution is defined as:

$$Correctness(o, o^{'}) = \begin{cases} 1 \; iff C(o) = C(o^{'}) \leftrightarrow P(o) = P(o^{'}) \\ 0 \; otherwise \end{cases} \tag{4.1}$$

Two objects are related perfectly when they are from the same class and if and only if they are assigned to the same cluster. In overlapping clustering, the multiplicity of observation occurrences

in classes and clusters is taken into account.

$$Multiplicity\ Precision(o, o^{'}) = \frac{Min(\left|P(o) \cap P(o^{'})\right|, \left|C(o) \cap C(o^{'})\right|)}{|P(o) \cap P(o^{'})|} \tag{4.2}$$

$$Multiplicity\ Recall(o, o^{'}) = \frac{Min(\left|P(o) \cap P(o^{'})\right|, \left|C(o) \cap C(o^{'})\right|)}{|C(o) \cap C(o^{'})|} \tag{4.3}$$

Multiplicity Precision is defined only if the observations o and $o^{'}$ have some cluster in common. Multiplicity recall is defined only if o and $o^{'}$ have some class in common. Therefore the relation between two observations can not be represented as a binary function. If two objects share one or more clusters, multiplicity precision is minimal when the two objects have no category in common. It is maximal if the number of common clusters is lower or equal of the number of common classes. Multiplicity precision increases if there is a matching class for each cluster. Multiplicity recall increases if there is a common cluster for a class shared by two objects. Analogously, multiplicity recall decreases if there are less shared clusters than shared classes while multiplicity precision decreases if there are less shared classes than clusters. The BCubed precision of an object represents the proportion of objects in its cluster having in common the same class with the object. The BCubed recall represents the proportion of objects in its class having in common the same cluster with the object. The overall BCubed precision is considered as the averaged precision of all the observations in the distribution.

$$Pecision\ BCubed = Avg_o \left[Avg_{o'.P(o)\cap P(o')\neq\phi} \left[Multiplicity\ precision(o, o^{'})\right]\right] \tag{4.4}$$

$$Recall\ BCubed = Avg_o \left[Avg_{o'.C(o)\cap C(o')\neq\phi} \left[Multiplicity\ recall(o, o^{'})\right]\right] \tag{4.5}$$

In the case where there is no overlap in the clustering, extended BCubed behaves as standard BCubed.

The F-measure represents the system efficiency in terms of recall and precision :

$$F_{measure} = \frac{1}{0.5 * \frac{1}{BCubed\ precision} + 0.5 * \frac{1}{BCubed\ recall}} \tag{4.6}$$

In order to focus the comparison on a unique evaluation measure, we have chosen the F-measure evaluation metric to build our comparison which represents the clustering system efficiency based on recall and precision.

Another significant measure for the overlapping clustering is the overlap size which is defined as:

$$Overlap\ size = \frac{1}{X} \sum_{x_i \in X} |A_i| \tag{4.7}$$

4.3 Experiments on real and artificial multi-labelled datasets

4.3.1 Data sets description

In order to evaluate the proposed methods in this work, the experiments were conducted on diverse domains for overlapping clustering researches where data requires to be assigned to more than one cluster:

- Artificial dataset: a generated dataset containing two classes, two attributes and 1000 observations.
- EachMovie dataset: A real data set used in video classification. Movies are grouped based on genres relying on user rating. EachMovie data set was collected when users entered numeric rating for different movies.
- Emotions dataset:Emotions data set represents music emotions classification domain, music songs are grouped according to emotions based on the analysis of music signals.
- Scene dataset: A real data set from image classification domain. It consists in grouping images according to landscapes based on the description of the spacial color moments.
- Yeast dataset: Yeast data set represents the domain of predicting the cellular localization sites of proteins.
- Iris dataset: Iris data set consists of 150 samples distributed on three species of Iris flower. The species are Iris Setosa, Iris Versicolor and Iris Virginia. The attributes characterizing Iris dataset samples are the petal length, The petal width, the sepal length and the sepal width. the attributes values are represented in centimetres. It is commonly used in pattern recognition literature as a test basis for a primary evaluation.

Table 4.1: Data sets description

Dataset	Attributes	Classes	Instances	Overlap
Artificial	2	1000	0.636	1
Iris	4	3	150	1
EachMovie	3	3	75	1.14
Emotions	72	6	593	1.81
Scene	300	6	2407	1.07
Yeast	117	14	2417	4.23

All the previously described datasets are multi-labelled in order to make it possible to use the BCubed evaluation method which is based on comparison between the provided labels and the clustering results.

4.3.2 Empirical results

We evaluate the quality of the clustering result using BCubed evaluation measures. Precision, recall, F-measure and overlap size are reported as averages of ten different initializations. All the methods were started with same initializations of cluster representatives in order to guarantee the same running conditions for all methods. For the methods offering overlap regulation, the overlap regulation parameter is adjusted in order to let the resulting clustering overlap rate match with the expected size of overlap according to each data set. The best obtained scores are represented with bold.

Table 4.2 shows that standard methods ALS_{lf2} and ALS_{lf1} produce more overlap than needed for the artificial dataset. When the overlap is regulated to match with the actual overlap of the data set, the clustering system performance increased. For example, F-measure increased form 0.627 with ALS_{lf2} to 0.908 with $RALS2$. We notice that for this data set, geometrical methods outperform additive methods as shown.

Table 4.2: Comparison of proposed methods with OKM, Parameterized ROKM, R_2OKM, ALS_{lf2} and ALS_{lf1} on the artificial dataset

		precision	Recall	F-measure	Overlap size
Geometrical	OKM	0.663	**0.991**	0.795	1.27
	Parameterized ROKM	**0.899**	0.919	0.909	**1**
	R_2OKM	0.892	0.936	**0.914**	**1**
Additive	ALS_{lf2}	0.482	0.896	0.627	1.31
	ALS_{lf1}	0.482	0.896	0.627	1.31
	$RALS2$	**0.908**	**0.909**	**0.908**	**1**
	$RALS1$	0.897	0.898	0.898	**1**

Analysing Table 4.3, we notice that the precision of regulated methods is higher than standard methods' precision. The F-measure have been improved with the use of overlap regulation. Recall of non regulated methods is the higher. which can be explained with the high overlap values produced by these methods, which does not match with the expected overlap for the Iris data set.

Table 4.3: Comparison of proposed methods with OKM, Parameterized ROKM, R_2OKM, ALS_{lf2} and ALS_{lf1} on the Iris dataset

		precision	Recall	F-measure	Overlap size
Geometrical	OKM	0.488	**0.994**	0.636	1.41
	Parameterized ROKM	**0.732**	0.841	0.786	**1**
	R_2OKM	0.722	0.853	**0.837**	**1**
Additive	ALS_{lf2}	0.376	**0.984**	0.531	1.65
	ALS_{lf1}	0.325	0.944	0.501	1.61
	$RALS2$	**0.761**	0.872	**0.765**	**1**
	$RALS1$	0.732	0.872	0.734	**1**

Table 4.4 shows that precision values increases with overlap regulated methods compared to non overlap regulated methods. We notice that recall values decreased with the proposed methods while the F-measure values are improved. For example, the comparison between the results of $RALS\,2$ and ALS_{lf2} shows that precision increased from 0.362 to 0.493 while recall decreased from 0.621 to 0.492 and F-measure value increased from 0.474 to 0.482. we also notice the size of overlap, $RALS\,2$ and $RALS\,1$ are able to generate precise overlap size according to the expected overlap of the data set while ALS_{lf2} and ALS_{lf1} generate clustering with higher overlap size than the actual data set overlap rate.

Table 4.4: Comparison of proposed methods with OKM, Parameterized ROKM, R_2OKM, ALS_{lf2} and ALS_{lf1} on the Emotions dataset

		precision	Recall	F-measure	Overlap size
Geometrical	OKM	0.387	**0.561**	0.426	2.12
	Parameterized ROKM	0.421	0.429	**0.443**	**1.81**
	R_2OKM	**0.431**	0.397	0.401	**1.81**
Additive	ALS_{lf2}	0.362	0.621	0.474	2.5
	ALS_{lf1}	0.333	**0.714**	0.47	2.55
	$RALS\,2$	**0.493**	0.492	0.482	**1.81**
	$RALS\,1$	0.489	0.546	**0.503**	**1.81**

Table 4.5 confirms the obtained results with the previous data sets. Precision increases with overlap regulated methods. For example the precision value obtained with $RALS\,2$ is 0.611 while the ALS_{lf2} precision value is 0.366. Recall decreases while F-measure is improved after the overlap regulation. Recall obtained with $RALS\,2$ is 0.721 and ALS_{lf2} recall is 894. Using $RALS\,2$, F-measure value 0.658 is improved compared to ALS_{lf2} F-measure value 0.519.

Table 4.5: Comparison of proposed methods with OKM, Parameterized ROKM, R_2OKM, ALS_{lf2} and ALS_{lf1} on the EachMovie dataset

		precision	Recall	F-measure	Overlap size
Geometrical	OKM	0.354	**0.806**	0.514	1.84
	Parameterized ROKM	0.716	0.718	0.717	**1.14**
	R_2OKM	**0.721**	0.720	**0.724**	**1.14**
Additive	ALS_{lf2}	0.366	**0.894**	0.519	1.6
	ALS_{lf1}	0.369	0.865	0.472	1.7
	$RALS2$	**0.611**	0.721	**0.658**	**1.14**
	$RALS1$	0.525	0.604	0.587	**1.14**

Table 4.6 shows the difference between overlap regulated methods and standard methods. Using ALS_{lf2} and ALS_{lf1} generate overlapping clusters with a high overlap size value respectively 2.6 and 2.65. The actual overlap size of the Scene data set is 1.07. We notice that non overlap regulated methods are unable to detect clusters with small overlap rate. the fact that $RALS2$ and $RALS1$ generate precise overlap size that matches with the actual overlap size of the data set, explains the large improvement in F-measure value. For example, F-measure value obtained using ALS_{lf1} is 0.184 while the F-measure value using $RALS1$ is 0.459.

Table 4.6: Comparison of proposed methods with OKM, Parameterized ROKM, R_2OKM, ALS_{lf2} and ALS_{lf1} on the Scene dataset

		precision	Recall	F-measure	Overlap size
Geometrical	OKM	0.124	**0.949**	0.219	2.51
	Parameterized ROKM	0.455	0.452	0.466	**1.07**
	R_2OKM	**0.490**	0.499	**0.494**	**1.07**
Additive	ALS_{lf2}	0.141	**0.911**	0.234	2.6
	ALS_{lf1}	0.103	0.907	0.184	2.65
	$RALS2$	0.391	0.417	0.412	**1.07**
	$RALS1$	b	0.453	**0.459**	**1.07**

Table 4.7 presents a comparison of proposed methods with existing methods on Yeast data set. This data set is characterized with a high computational level because of its important number of classes and its high number of attributes and observations. We notice that for this data set, OKM can't reach the required overlap size. Which means hat it produces a clustering that lacks information. However ALS_{lf2} generates a clustering that largely overpasses the needed overlap size which is not an appropriate classification. $RALS2$ shows an important improvement in the clustering efficiency, i.e. the F-measure have been widely improved from 0.234 to 0.575. This is due to the overlap regulation which restricted the overlaps to match with the actual overlap of Yeast data set. This makes precision value increase which affects the f-measure value. Yet, we couldn't perform the clustering process with ALS_{lf1} on Yeast data set. The reason is that ALS_{lf1} is characterized by a high complexity algorithm.

Table 4.7: Comparison of proposed methods with OKM, Parameterized ROKM, R_2OKM, ALS_{lf2} and ALS_{lf1} on the Yeast dataset

		precision	Recall	F-measure	Overlap size
Geometrical	OKM	**0.669**	0.381	0.486	3.40
	Parameterized ROKM	0.600	**0.504**	0.548	**4.23**
	R_2OKM	0.620	0.500	**0.565**	**4.23**
Additive	ALS_{lf2}	0.141	**0.911**	0.234	7.9
	ALS_{lf1}	-	-	-	-
	$RALS2$	**0.643**	0.441	**0.575**	**4.23**
	$RALS1$	-	-	-	-

4.4 Discussion

For all experiments, the overlap size influences the value of obtained F-measure. The value of precision decreases and the value of recall increases as the overlap size increases. OKM, ALS_{lf2} and ALS_{lf1} generate large overlaps compared with the actual overlap size of each data set. For example, ALS_{lf2} applied to EachMovie data set, Emotions data set and Scene data set generates clusterings with respectively overlaps sizes of 2.5, 1.6 and 2.6. The actual overlap size for each data set is respectively 1.14, 1.81 and 1.07.

Using overlap regulation on non overlap regulated methods (ALS_{lf2} and ALS_{lf1}) by adjusting the overlap regulation parameter reduces the overlap size. For example, $RALS2$ used on Scene data set reduces the overlap size from 2.6 to 1.07. We notice that non overlap regulated methods have the best values of recall. The reason is that those methods build clusters with large overlapping boundaries. Overlap regulated methods have the best precision values. The reason is that they build clusters with smaller overlap rates than non regulated overlap methods. Comparing overall F-measure values obtained by proposed methods outperforms F-measure values obtained by ALS_{lf2} and ALS_{lf1}. This improvement is due to the reasonable overlap sizes compared with the real overlap size of each data set.

For the proposed methods $RALS2$ and $RALS1$ compared respectively to ALS_{lf2} and ALS_{lf1}, we

notice the performance of these methods to detect clusters with the required overlaps. $RALS\,2$ and $RALS\,1$ are able to detect small and large overlaps depending to the overlap regulation parameter α. The improvement in overall F-measure is important. For example, in Scene data set the F-measure obtained with $RALS\,1$ is 0.459 while using ALS_{lf1} we obtain 0.184. This improvement is induced by the improvement of precision. The overall precisions obtained with $RALS\,2$ and $RALS\,1$ are characterized with by high values compared respectively to ALS_{lf2} and ALS_{lf1}. The restriction of assignments explains the precision improvement.
Figure 4.1 shows the size of overlap generated by each methods for every data set. Non overlap regulated methods generate higher overlap size than the expected overlap except for Emotions data. in Emotions data set, the overlap size is 1.81. OKM generates 1.84 overlap size while ALS_{lf2} generates 1.6 and ALS_{lf1} produces 1.7 overlap rate. However, in four out of five cases, non overlap regulated methods highly overpass the expected size of overlap. This problem is solved with overlap regulated methods which produce the expected size of overlap. The overlap manipulation helps to increase the F-measure which have an influence on the clustering system efficiency.

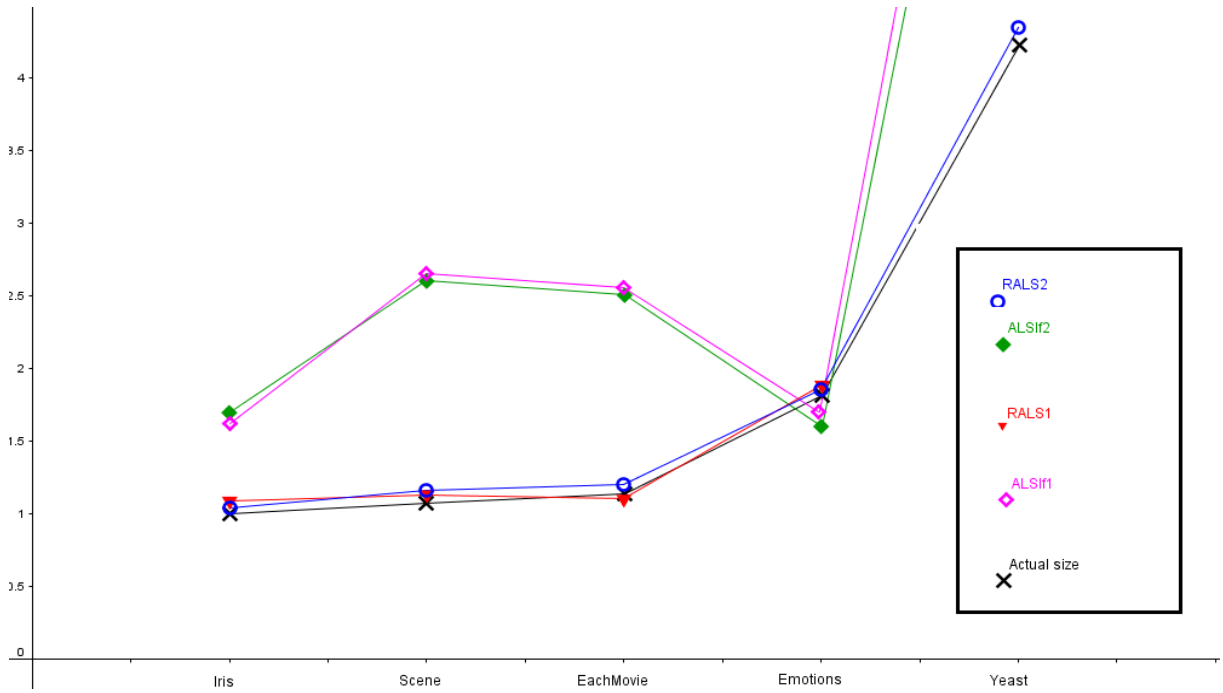


Figure 4.1: Overlap size produced by every method on each data set

Figure 4.2 and figure 4.3 shows how the F-measure is improved with the introduction of

overlap regulation in ALS_{lf2} and ALS_{lf2}. For all the data sets, $RALS\,2$ and $RALS\,1$ produces a clustering characterized with higher value of F-measure. This implies that the system efficiency with $RALS\,2$ and $RALS\,1$ is respectively better than with ALS_{lf2} and ALS_{lf1}.

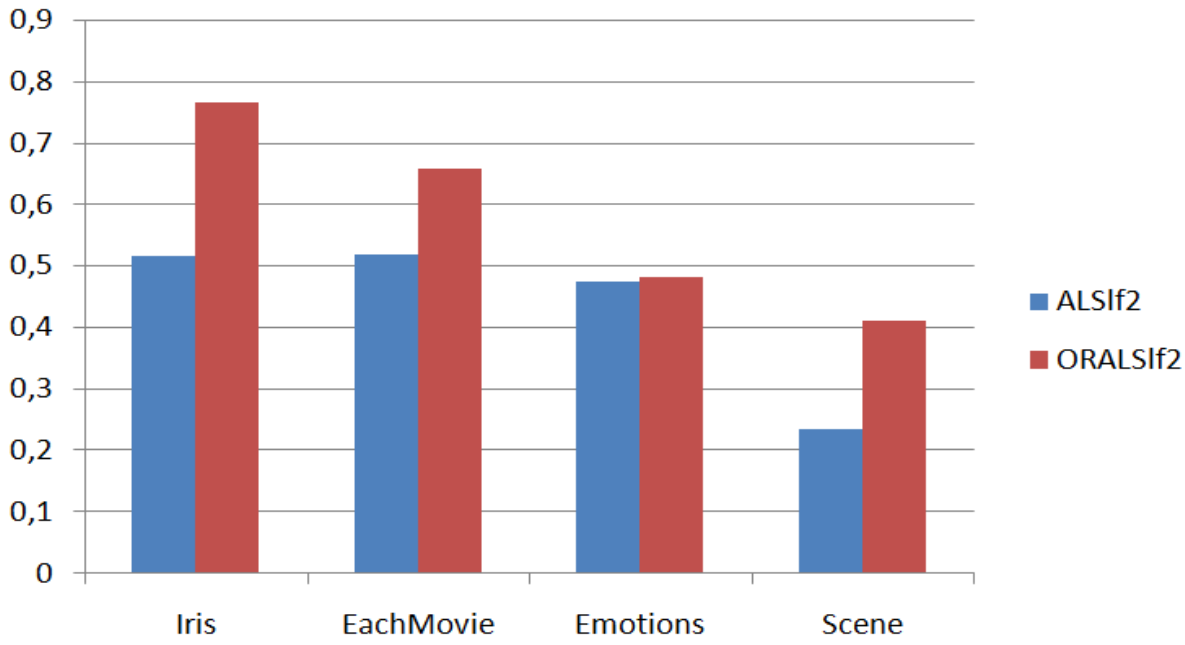


Figure 4.2: F-measure improvement with the use of overlap regulation on ALS_{lf2}

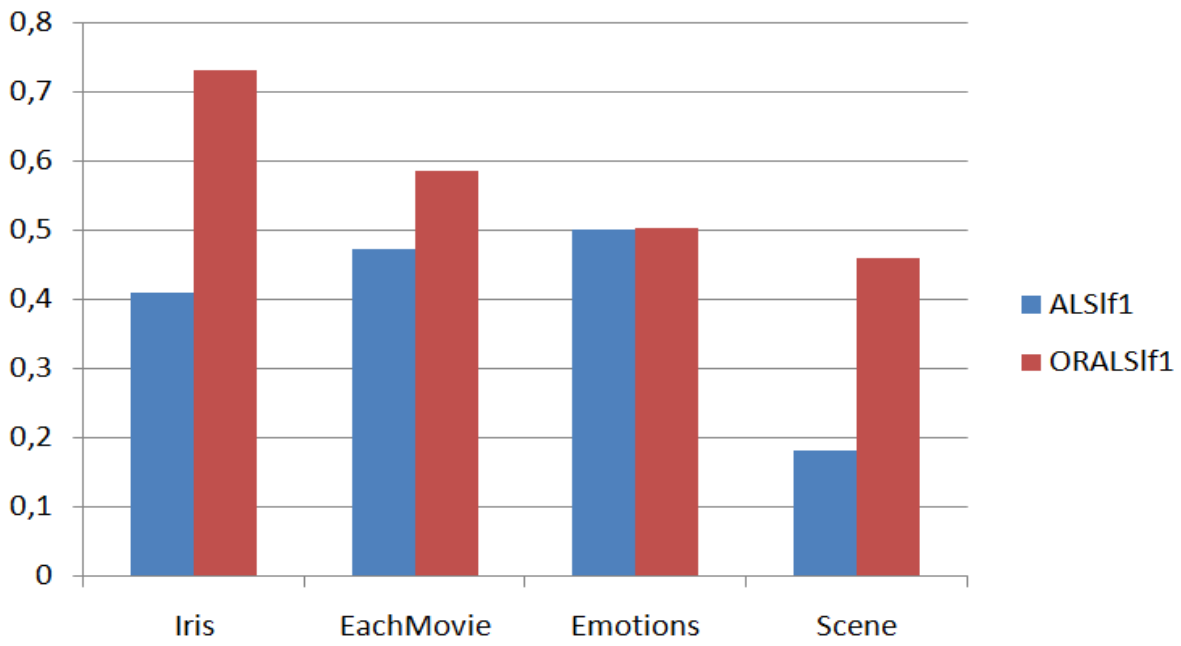


Figure 4.3: F-measure improvement with the use of overlap regulation on ALS_{lf2}

$RALS2$ and $RALS1$ methods respectively outperform ALS_{lf2} and ALS_{lf1} as shown on Histograms 4.2 and 4.3. Anyway, in the majority of the cases, geometrical overlap regulated methods $ParameterizedROKM$ and $R2OKM$ show better results than additive overlap regulated methods. For example, on EachMovie, Iris, Scene datasets, $ParameterizedROKM$ and $R2OKM$ show higher values of F-measure than both $RALS2$ and $RALS1$. For the Emotions data set, Additive overlap regulated methods show better results and higher F-measure values.

4.5 Conclusion

In this chapter, we have detailed the evaluation methodology that we used in order to evaluate the performance of the proposed methods, . We have described the datasets on which the experiments were performed. Empirical results showed the efficiency of $RALS1$ and $RALS2$ in detecting clusters by regulating the size of the expected overlap. The comparison of $RALS1$ and $RALS2$ to existing partitional overlapping clustering methods showed the important improvement of the generated clustering's quality. We have also made a comparison between additive and geometrical methods.

Conclusion

As a machine learning technique, clustering remains well commonly used in the data mining domain. In this work, we focus on the partitional overlapping clustering which takes into consideration the possibility to assign observations simultaneously to several groups. Many solutions were proposed in the purpose to deal with this issue additive and geometrical methods. Additive methods consider the overlap as the profiles addition of the related clusters (ALS_{lf1}, ALS_{lf2}) while geometrical methods consider the overlap as the profiles mean of the related clusters(OKM). Despite the effectiveness of these methods, they fail to accomplish the clustering task in an appropriate way since they produce a fixed overlap size which is practically different from the expected overlap size that allows to fit the concerned data structure. This problem was only solved with geometrical model and reminds unsolved for the additive model. Therefore, we propose to regulate the overlap the size of overlaps generated by additive methods in the purpose to detect an appropriate classification when an overlapping clustering is expected. The proposed methods are referred to as $RALS\,1$ and $RALS\,2$ introducing an overlap regulation principle based on the number of assignments. The different experiments which we have performed on real and artificial data sets showed the important improvement in the efficiency of additive overlapping clustering methods. In fact, with the proposed methods the expert have to try different values of the overlap regulation parameter in order to obtain the required clustering. We plan to propose a methods that auto-adjusts the regulation parameter in order to offer an automatic overlap regulation instead of the manual regulation. On the other, the number of clusters K represents also an issue. Usually, the user executes the clustering process several times with different numbers of clusters in order to determine the expected number of clusters. We aim to propose an additive method that auto-detects the number of clusters from the dataset. We also aim to study different distance metrics in the purpose to see the behaviour of the methods with different distances which

will help to determine the distance metric that best fits with the additive methods. we also intend to propose overlap regulated methods based on the dispersal of the cluster. we plan to analyse the different parameters that affects the clustering result quality such as the number of clusters, the distance metric, the input parameters and the data characteristics. In the literature, there is no much known techniques that helps in determining the input parameters values in order to best fit the data structure. We aim in the future to propose a method that makes it possible to find the optimal combination of parameters that produces the best clustering result.

References

Ackermann, M. R. (2009). *Algorithms for the bregman k-median problem* (Unpublished doctoral dissertation). Paderborn, Univ., Diss., 2009.

Amigó, E., Gonzalo, J., Artiles, J., & Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*, *12*(4), 461–486.

Arabie, P., & Carroll, J. D. (1980). Mapclus: A mathematical programming approach to fitting the adclus model. *Psychometrika*, *45*(2), 211–235.

Bagga, A., & Baldwin, B. (1998). Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics-volume 1* (pp. 79–85).

Banerjee, A., Merugu, S., Dhillon, I. S., & Ghosh, J. (2005). Clustering with bregman divergences. *The Journal of Machine Learning Research*, *6*, 1705–1749.

Battle, A., Segal, E., & Koller, D. (2005). Probabilistic discovery of overlapping cellular processes and their regulation. *Journal of Computational Biology*, *12*(7), 909–927.

Ben NCir, C.-E., Cleuziou, G., & Essoussi, N. (2014). Generalization of c-means for identifying non-disjoint clusters with overlap regulation. *Pattern Recognition Letters*, *45*, 92–98.

Ben N'Cir, C.-E., Essoussi, N., & Limam, M. (2015). Kernel-based methods to identify overlapping clusters with linear and nonlinear boundaries. *Journal of Classification*, *32*(2), 176–211.

Ben NCir, C.-E. B., Cleuziou, G., & Essoussi, N. (2015). Overview of overlapping partitional clustering methods. In *Partitional clustering algorithms* (pp. 245–275). Springer.

Black, P. E. (2006). Manhattan distance. *Dictionary of Algorithms and Data Structures*, *18*, 2012.

Bonchi, F., Gionis, A., & Ukkonen, A. (2011). Overlapping correlation clustering. In *Data mining (icdm), 2011 ieee 11th international conference on* (pp. 51–60).

Cantrell, C. D. (2000). *Modern mathematical methods for physicists and engineers*. Cambridge University Press.

Carroll, J. D., & Arabie, P. (1983). Indclus: An individual differences generalization of the adclus model and the mapclus algorithm. *Psychometrika*, *48*(2), 157–169.

Chaturvedi, A., & Carroll, J. D. (1994). An alternating combinatorial optimization approach to fitting the indclus and generalized indclus models. *Journal of Classification*, *11*(2), 155–170.

Cherkassky, V., & Mulier, F. (1998). Learning from data: Concepts, theory, and methods (adaptive and learning systems for signal processing, communications and control series).

Cleuziou, G. (2007). Okm: une extension des k-moyennes pour la recherche de classes recouvrantes..

Cleuziou, G. (2008). An extended version of the k-means method for overlapping clustering. In *Pattern recognition, 2008. icpr 2008. 19th international conference on* (pp. 1–4).

Cleuziou, G. (2009). Okmed et wokm: deux variantes de okm pour la classification recouvrante. In *Egc* (pp. 31–42).

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

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, *50*(1), 1–18.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

Depril, D., & Van Mechelen, I. (2005). One-mode additive clustering of multiway data..

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

Depril, D., Van Mechelen, I., & Wilderjans, T. F. (2012). Lowdimensional additive overlapping clustering. *Journal of classification*, *29*(3), 297–320.

Diday, E. (1987). Orders and overlapping clusters by pyramids.

Duda, H., Hart, P., et al. (2001). *Stork, pattern classification.* John Wiley & Sons.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise..

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

Fraley, C., & Raftery, A. E. (2003). Enhanced model-based clustering, density estimation, and discriminant analysis software: Mclust. *Journal of Classification*, *20*(2), 263–286.

Ganguly, A. R., Gama, J., Omitaomu, O. A., Gaber, M., & Vatsavai, R. R. (2008). *Knowledge discovery from sensor data*. CRC Press.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques: concepts and techniques*. Elsevier.

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

Jain, A. K., Duin, R. P., & Mao, J. (2000). Statistical pattern recognition: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *22*(1), 4–37.

Kaufman, L., & Rousseeuw, P. J. (n.d.). Clustering large applications (program clara). *Finding groups in data: an introduction to cluster analysis*, 126–163.

Kaufman, L., & Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, 68–125.

Khachumov, M. (2012). Distances, metrics and cluster analysis. *Scientific and Technical Information Processing*, *39*(6), 310–316.

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

Lesot, M.-J., Rifqi, M., & Benhadda, H. (2008). Similarity measures for binary and numerical

data: a survey. *International Journal of Knowledge Engineering and Soft Data Paradigms*, *1*(1), 63–84.

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

MacQueen, J., et al. (1967). Some methods for classification and analysis of multivariate observations..

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

McJones, P. (1997). *Eachmovie collaborative filtering data set, dec systems research center, 1997.*

Mirkin, B. (1987). The method of principal clusters. *Automation and remote control*, *48*(10), 1379–1388.

MIRKIN, B. (1990). Cluster-analysis on the basis of a linear-model-generalizing article [CITATION]. *INDUSTRIAL LABORATORY*.

Ng, R. T., & Han, J. (2002). Clarans: A method for clustering objects for spatial data mining. *Knowledge and Data Engineering, IEEE Transactions on*, *14*(5), 1003–1016.

Oja, E. (1992). Principal components, minor components, and linear neural networks. *Neural Networks*, *5*(6), 927–935.

Schikuta, E., & Erhart, M. (1997). The bang-clustering system: Grid-based data analysis. In *Advances in intelligent data analysis reasoning about data* (pp. 513–524). Springer.

Sheikholeslami, G., Chatterjee, S., & Zhang, A. (2000). Wavecluster: a wavelet-based clustering approach for spatial data in very large databases. *The VLDB Journal*, *8*(3-4), 289–304.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*(2), 87.

Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319–2323.

Trohidis, K. (2008). Multi-label classification of music into emotions..

Xu, R., Wunsch, D., et al. (2005). Survey of clustering algorithms. *Neural Networks, IEEE*

Transactions on, *16*(3), 645–678.

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

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). Birch: an efficient data clustering method for very large databases. In *Acm sigmod record* (Vol. 25, pp. 103–114).