U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

PhD

3.º CICLO

FCUP
2015

U.PORTO

Contribution to the knowledge of hierarchical clustering algorithms and consensus clustering
Studies applied to personal recognition by hands biometrics

Lúcia de Paiva Martins de Sousa

FC

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Contribution to the knowledge of hierarchical clustering algorithms and consensus clustering

Studies applied to personal recognition by hands biometrics

Lúcia de Paiva Martins de Sousa

Tese de Doutoramento apresentada à
Faculdade de Ciências da Universidade do Porto,
Departamento de Matemática

2015

PhD

Contribution to the knowledge of hierarchical clustering algorithms and consensus clustering

Studies applied to personal recognition by hands biometrics

Lúcia de Paiva Martins de Sousa
Programa Doutoral em Matemática Aplicada
2015

Tese submetida à Faculdade de Ciências da Universidade do Porto
para a obtenção do grau de Doutor em Matemática Aplicada

Orientador
João Gama, Professor Associado, Faculdade de Economia da Universidade do Porto

Coorientador
Katti Faceli, Professora Adjunta, Departamento de Ciências dos Computadores da Universidade Federal de São Carlos

Resumo

Na análise exploratória de dados, os algoritmos de agrupamentos hierárquicos com suas características podem fornecer diferentes agrupamentos quando aplicados ao mesmo conjunto de dados. Na presença de vários agrupamentos, cada um identificando uma específica estrutura dos dados, os consensos de agrupamentos fornecem uma contribuição para lidar com essa questão.

Este trabalho é composto por duas partes:

Na primeira parte, propomos explorar o perfil dos agrupamentos hierárquicos de base, em função das suas variabilidades, para obtenção do consenso de agrupamentos. Como um primeiro resultado das nossas pesquisas, identificamos a técnica de consenso com melhor desempenho que as demais, em função das características dos agrupamentos hierárquicos usados como iniciais. Este resultado permite-nos identificar uma condição suficiente para a existência de consenso de agrupamentos, assim como também definir uma nova estratégia para avaliar os consensos. Permite-nos ainda o estudo de uma nova propriedade dos algoritmos de agrupamentos hierárquicos.

Na segunda parte, exploramos uma aplicação do mundo real. Numa primeira análise, usamos conjuntos de dados biométricos extraídos pelas mãos para reconhecimento pessoal. Mostramos que os agrupamentos hierárquicos, obtidos pelo

algoritmo SEP/COP, podem fornecer resultados com grande precisão quando aplicável a esses conjuntos de dados. Além disso, descobrimos que é possível um reconhecimento de 100% mais do que na literatura. Numa segunda análise, consideramos a aplicação das técnicas de consensos de agrupamentos ao problema da identificação da parentalidade de pessoas pelas biometrias das mãos. Os resultados que obtivemos indicam que a fotografia da mão tem informação que permite a identificação dos familiares de pessoas mas, em concreto nos nossos dados não obtivemos resultados muito positivos (observamos uma probabilidade de 95% de o pai ou a mãe e 94% de um irmão estar na metade das mãos mais parecidas) que, pensamos, estar ligado à fraca qualidade das fotografias que usamos. Mas os resultados indicam que a técnica tem potencial e que, se a recolha das fotografias for feita num *scanner* com pinos fixos, a mão pode ser uma alternativa interessante na identificação da parentalidade de crianças perdidas aquando da aplicação dos consensos de agrupamentos.

Abstract

In exploratory data analysis, hierarchical clustering algorithms with its features can provide different clusterings when applied to the same data set. In the presence of several clusterings, each one identifying a specific data structure, consensus clustering provide a contribution to deal with this issue.

The work reported here is composed by two parts:

In the first part, we intend to explore the profile of base hierarchical clusterings, according to their variabilities, to obtain the consensus clustering. As a first result of our researches, we identified the consensus clustering technique as having better performance than the others, depending on the characteristics of hierarchical clusterings used as base. This result allows us to identify a sufficient condition for the existence of consensus clustering, as well as define a new strategy to evaluate the consensus clustering. It also leads to study a new property of hierarchical clustering algorithms.

In the second part, we explore a real-world application. In a first analysis, we use data sets derived by biometrics extracted from hands for personal recognition. We show that the hierarchical clusterings obtained by SEP/COP algorithms, can provide results with great accuracy when applied to these data sets. Furthermore, we found an increased 100% of recognition rate, comparing to the ones found in literature. In a second analysis, we consider the application of consensus clustering techniques to the

problem of the identification of people's parenting by the hands biometrics. The results obtained indicate that hand’s photography has information that allows the identification of people’s family members but, according to our data, we didn't have very positive results (we observed a probability of 95% of the parents, and 94% of a sibling to be in the half of the more similar hands) that we believe it’s due to the poor quality of the photographs we used. However, the results indicate that the technique has potential, and if the collection of photographs is made using a scanner with fixed pins, the hand may be an interesting alternative for the identification of parenting of missing children when it is applied the consensus clustering.

Acknowledgments

I would like to thank my supervisors, João Gama and Katti Faceli for the great contribution for this work. I would also like to thank Professor Pedro Cosme for the suggestions and encouragement constantly given over the years and a very special thanks to my family by understanding my absence in recent years.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The subjects handled herein insert up on Data Mining area, for the need to operate data sets derived from several subjects.

The need to efficiently treat information extracted from data sets increased the interest on developing effective tools for its organization. Machine Learning, as Data Mining is the area that addresses this subject, learning from the data.

Regarding the exploratory data analysis, Data Mining allows analysing data sets discovering and extracting interesting patterns such as clusters.

Clustering is one of the most important unsupervised learning tools when no prior knowledge about the data set is available. Clustering algorithms aim to find the underlying structure of the data sets considering clustering criteria, properties in the data and specific way of data comparison [97]. In literature many clustering algorithms have been proposed as having a common goal which is, given a set of objects, to

group them into clusters, in such way that similar objects are in the same cluster and dissimilar objects are in different clusters.

The hierarchical clustering algorithms provide several clustering structures which are represented by a hierarchy. A hierarchy allows an easy user interaction and at the same time detect different clusterings which may lead to the discovery of unknown underlying patterns of the data set. These algorithms applied to a data set, always provide a hierarchy even when the data set is completely random, i.e., absent of cluster structure. So, it is necessary to consider the results as proposals to validate.

Considering a data set with a cluster structure, it's known that different hierarchical clustering algorithms, with its own characteristics and criteria, can provide different cluster structures when applied to this data set. Also, there is no single algorithm capable of matching all possible cluster structures, according to the number and shape of the clusters [38].

These problems introduce a concern that can lead to searching validation processes. The implementation of measures of clustering validity arises as a contribution to aid their interpretation and decision making. Several techniques of clustering validation emerged in literature, applying statistical measures, to help selecting the most appropriate algorithm. Most of the validation measures are biased, each one favouring a different clustering criterion. Moreover, they lead to the selection of one single best clustering, among various relevant structures that can be hidden in the data, thus limiting the discovery of new knowledge [26].

Trying to overcome the issues mentioned above, rather than selecting one single clustering among the various, some researchers have decided to combine them. Clustering combination or consensus clustering is a technique that combines information of multiple clusterings obtained from the same data set, providing a consensus solution. This has proven to be a better alternative than the single clustering [31].

Over the past years, several consensus clustering techniques emerged considering each one implicit assumptions and a specific way of providing the consensus solution. So, different consensus clustering techniques can find different

consensus clustering. Some investigations with the goal of improving the consensus solutions impose that clusterings to combine must have different weight by the fact that these clusterings may not have the same quality [96].

As the clustering algorithms provide clusterings even in absence of a cluster structure in data sets, the consensus techniques always provide a consensus clustering even facing the possibility of having no consensus.

These difficulties concerning to consensus clustering algorithms constitute the first motivation of this thesis.

Another motivation arises from the application of clustering analyses to the real-world data set derived from the hands biometrics for recognition.

Recognition systems based on hand biometry are very popular and are among the oldest biometric tools used for automatic person authentication. Devices for controlling access based on these systems have been manufactured and marketed since the late 70's, and used, for example, in airports [69].

Researches in the field of biometrics found that the human hand contains features that can be used for personal identification, such as, geometry and shape of the hands [30]. A biometric system of hand recognition extracts the most relevant features of the hand and with these the signature of the correspondent person is created. Usually, this signature represents the identity of the person in a system that is used for recognition by comparing it with the existing set of features in the database [69].

Hand biometrics recognition systems, as well as the applied technologies, have been developed in recent decades. These systems comprise several steps, since images acquisition to features extraction, including the construction of the database with the peoples signatures, and at last, the recognition. Different systems apply different commitments relative to each step.

Many of these systems arise in literature having a common idea which is mainly achieving 100% rate of people identification using large databases with the people

signatures. This is a problem that most of the systems are not able to concretize and the motivation to our researches.

1.1 Thesis goals and proposed solutions

This thesis explores the competence of the hierarchical clustering algorithms and the consensus clustering techniques. Empirical studies are performed to achieve the proposed objectives.

In a first study, the goals of this thesis are:

- Find conditions for the existence of consensus clustering;
- Propose a new strategy to evaluate the consensus clustering.

The studies of the hierarchical clustering algorithms regarding their variability allow to define clusterings profiles able to give solution to both of the mentioned goals. Clusterings, derived by an algorithm, with great variability between them, constitute the base clusterings able to lead to a consensus clustering and moreover they lead to the best consensus clustering.

Also, these studies about the clustering algorithms variability contribute to the study of a new property of the hierarchical clustering algorithms. Each hierarchical clustering algorithm is better suited to be applied to data sets with certain characteristics of clusters. Applying an algorithm better suited to a data set, this algorithm presents stability by the low variability obtained.

These contributions are included in a paper's version submitted to an international journal [85].

The other goals, in another study, are related to the real-world application, the recognition by hand's biometrics.

First, we intend to explore the potential of the hierarchical clustering algorithms. For that, we apply the usual hierarchical clustering algorithms and a different approach, SEP/COP [37]. This approach consists in a different interpretation of the hierarchy. We aim to apply both algorithms to the problem of hand recognition by biometrics, namely on the recognition stage. We discover that the SEP/COP algorithms, can achieve a great performance including persons recognition by hands biometrics, reaching 100% of recognition. Furthermore, our results outperform the results in literature. This contribution is published in [84]. Also, a preliminary version of this work was presented in [82].

Secondly, considering the great potential of hands biometrics, we propose the consensus clustering techniques to cope with the challenges of parental recognition.

There are many studies addressing recognition by hands biometry but as far as we investigated, there is no study in literature addressing the problem of person parenthood identification, based in hands biometrics. So, this is a challenge of great importance which is framed in this thesis.

Applying the consensus clustering techniques to children and parents hands biometrics, no consensus achieves a great performance. Despite this, we discovered that it is possible to find a person's parents by restricting the searched database. This contribution is included in a paper's version submitted to an international journal [86]. Also, a preliminary version of this work was presented in [83].

1.2 Thesis organization

There are two considered parts on this thesis organization. The first one is formed by the Chapters 2, 3 and 4, which support into the first goals of this thesis. At the second part, formed by the remaining Chapters 5, 6 and 7, we proceed to apply the algorithms studied to hands biometrics recognition problem. More specifically:

In Chapter 2, we review the properties and characteristics own of some hierarchical clustering algorithms. As well as, some consensus clustering approaches and their procedures to obtain the consensus, considering the approaches most referred in literature. Other approaches to the consensus clustering are briefly presented.

Chapter 3 and Chapter 4 present some of the contributions to this thesis.

In Chapter 3, in context of clustering validation, we start by presenting some works dedicated to this subject, as well as some measures to evaluate the variability of the clustering algorithms. After, we study the variability of the hierarchical clustering algorithms which allows to define profiles of the base clusterings for the obtainment of consensus clustering. These researches lead to the fulfilment of the first goals of this thesis. The obtainment of the consensus clustering is performed in Chapter 4.

In Chapter 4, addressing the validation of the consensus result, we proceed to analyse the performance of the consensus techniques considering the variability of the base clusterings.

Chapter 5 addresses the real world application of hands biometrics for recognition that is applied in Chapters 6 and 7. We review in literature several contributions to this subject that have been emerging over the years.

Chapter 6 inquires the potential of the hierarchical clustering algorithms as applied to the person's recognition by the hands biometrics. It is presented an approach considering a different interpretation of the hierarchy produced by the

hierarchical clustering algorithms, the SEP/COP algorithms. It is provided a comparison of the performances between the usual hierarchical clustering algorithms and the approach SEP/COP. These studies lead to another contribution to this thesis.

Chapter 7 explores the potential of the consensus clustering techniques. The traditional consensus clustering techniques (studied in Chapter 2) and the multi-objective MOCLE are analysed. We propose to investigate if it is possible to recognize a person's parents by the hands biometrics and by applying these techniques. We describe the procedures to construct our database which contains hands images of parents and children. These studies contribute to an innovative work in applications related to the parental recognition by the hands biometry.

Chapter 8 provides the final conclusions of this work and the future works.

Chapter 2

Hierarchical clustering algorithms and consensus clustering

2.1 Summary

This Chapter is addressed to the hierarchical clustering algorithms regarding their methodology of aggregating clusters and their known characteristics. Also, it is addressed the consensus clustering approaches and some traditional consensus clustering techniques most referred in literature. Comparisons between some techniques referenced in literature are presented.

2.2 Introduction to the hierarchical clustering algorithms

The clustering algorithms are much applied in Data Mining, and widely used to solve real problems from various fields such as Medicine, Psychology, Botany, Sociology, Biology, Archeology, Marketing, etc. [65]. They are unsupervised learning algorithms aiming to find a clustering of a given data set, such that, similar elements are in the same cluster and distinct elements belong to different clusters. Among various clustering algorithms, the hierarchical clustering algorithms are oftentimes applied, owing their easy implementation and inherent advantages to the graphical representation of the resultant partitions, through a dendrogram.

The clustering algorithms can be classified into two main categories, as, hierarchical and partitional. The partitional algorithms generate a single data partition, while hierarchical algorithms organize the data into a nested sequence of partitions [46].

A hierarchical clustering method generates a hierarchy that is a structure with more information than the clustering obtained by partitional algorithms. Moreover, it doesn't need to specify the numbers of clusters, and most of the hierarchical clustering algorithms are deterministic. In addition to these advantages, the hierarchical clustering algorithms have lower cost than the traditional algorithms such as K-means or EM (Expectation Maximization), but instead, do not scale well and have, at least, time complexity of $O(n^2)$, where n is the number of objects [68], [32].

Hierarchical clustering algorithms produce a set of nested clusters organized in a hierarchy, represented in a dendrogram. These algorithms can be, divisive (top-down) or agglomerative (bottom-up). An agglomerative algorithm considers, at first, each element of the data set as a cluster, and then successively joins pairs of clusters until all clusters are combined into a single cluster containing all the elements. A divisive clustering algorithm starts with a cluster with all elements and then divides the clusters recursively until obtain clusters with the individual elements [68]. Because the agglomerative algorithms are most often used than the divisive ones, this work addresses these algorithms, and henceforth we refer only to these algorithms.

A hierarchical algorithm constitute a methodology of sequentially aggregate, not just pairs of clusters, but also can join two elements or objects forming a new cluster, or still, add an element to an existing cluster. Initially, each element forms a cluster. The process is carried out by ordered steps of aggregation where the order of each step corresponds to the level of the hierarchy. These aggregations are based on proximities or similarities matrix, which represent the distance between elements and (or) clusters. The idea is to observe the proximity matrix (or a representation in graph), and in accordance with the shortest distance, joins the elements in a cluster and (or) join the corresponding clusters, thus building a new cluster. With the appearance of a new cluster, distances are recalculated and thus, one gets a new proximity matrix. The process ends when all elements are at the same cluster. The final result is a hierarchy of partitions represented in a dendrogram. Analysing the dendrogram, one can cut it in different levels, by a horizontal line, yielding different partitions or clustering with different number of clusters. At our studies, we decided to fix the cut level, i.e., fix the number of clusters according the data sets and their known structure.

The various aggregation methods differ in how they define the distance between clusters, i.e., differ in the entries of the proximity matrix. Different definitions of the distances result in different clustering methods [46].

The distance between two clusters, C_1 and C_2, are stated by distance between objects. Given an object $x = (x_1, x_2, \dots, x_{\mathrm{d}})$, where d, is the dimensionality of the data set, the distance between two objects can be calculate by different metrics such as:

• Euclidian- $d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$, (2.1)

• Manhattan- $d(x,y) = \sum_i |x_i - y_i|$, (2.2)

• Maximum- $d(x,y) = \max_i |x_i - y_i|$, (2.3)

• Mahalanobis- $d(x,y) = \sqrt{(x-y)^T S^{-1}(x-y)}$, where S is the covariance matrix [95]. (2.4)

At our experiments, we apply the Euclidian distance. This metric corresponds to the trivial sense of distance and it's the most known and used metric [46]. Also, in our preliminary experiments, this metric, was found to be preferable compared to the Mahalanobis metric. As it takes into consideration the correlation between the data sets, the covariance matrices can be difficult to determine and memory and computation time grows in a quadratic way with the number of features [64].

The hierarchical clustering algorithms have different ways to define $d(C_1, C_2)$, for instance:

- In Single-Linkage method (SL), it is the distance between pair of elements (one in each cluster), which are the closest among all possible pairs,

$$d(C_1, C_2) = \min_{x \in C_1, y \in C_2} d(x, y). \quad (2.5)$$

- In Complete-Linkage method (CL), it is the distance between pair of elements (one in each cluster), which are the most distant from all possible pairs,

$$d(C_1, C_2) = \max_{x \in C_1, y \in C_2} d(x, y). \quad (2.6)$$

- In Average-Linkage method (AL), it is the average distance between all pairs of elements (one in each cluster),

$$d(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{x \in C_1} \sum_{y \in C_2} d(x, y). \quad (2.7)$$

Ward's method (W), also known as the method of minimum variance, differs from the above mentioned methods, not using distances between clusters to aggregate them. The objective of W is to look for the slightest deviation between the cluster centroid and the other elements of the cluster, i.e., looking for the smallest variance of the cluster. At each step all the possibilities of adding two clusters are checked, and is chosen the one which causes the smallest increase of the sum of squares error, SSE, of the aggregate cluster. Being,

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i.}\right)^2, \quad (2.8)$$

where k is the number of clusters, y_{ij} the j^{th} element in the i^{th} cluster which has centroid $\bar{y}_{i.}$ and n_i elements.

Due to the characterization of the similarity between pairs of clusters that these methods do, they often provide different hierarchies and therefore, different partitions, for the same data set. For instance, SL establishes a local aggregation strategy, i.e., takes into account only the area where two clusters are closer to one another. The other parts of clusters, as well as the general structure of the clustering are not taken into account. So, the clusters produced by SL are not compact and tend to be elongated [68]. On the other hand, CL avoids this chain effect problem, the aggregation of clusters is not local, and the whole structure of the clustering can affect the decisions of aggregation. CL produces compact clusters with approximately the same size (number of elements) and smaller diameters. It is also sensitive to outliers. A single element far from the center can, dramatically increase the diameters of candidate clusters to join together and completely change the final clustering [68]. SL is more versatile than CL and works well in data sets containing non-isotropic clusters, including clusters well separated and concentric, while CL works well in data sets with clusters that may not be well separated [46]. The drawbacks of SL and CL are due to the way they calculate the similarity between clusters by the similarity of a single pair of elements. AL, otherwise, evaluates similarities between clusters based on all their elements. Thus, AL overcomes the sensitivity of CL to outliers and the performance of SL forming long chains that do not correspond to the intuitive notion of compact clusters with spherical shapes [68]. On the other hand, W, intending to minimize the variance of the cluster's elements favors compactness of the clusters. The distance between two clusters is defined as the consequent increase in SSE if both clusters would join to form a single one. W has better performance than other hierarchical methods, specially, when the clusters proportions are approximately equals [29]. Some principal characteristics of SL, CL, AL and W algorithms are summarized in Table 2.1.

Table 2.1: Main properties of SL, CL, AL and W algorithms.

SL [65], [90]	CL [34], [46], [90]	AL [68]	W [3], [4], [29]
Favors connectivity of the clusters	Favors compactness of the clusters	Clusters tend to spherical shapes	Favors compactness of the clusters
Detect clusters with arbitrary shapes and the same density	Imposes clusters with spherical shapes	Is less susceptible to noise and outliers than CL and SL	Tends to create clusters with the same number of elements and few elements
Does not deal well with different densities clusters	Tends to divide large clusters		Is slightly sensitive to outliers and noise
Produces large, elongated and well separated clusters	Produces small clusters, more balanced (with same diameter) and closest		
Is sensitive to outliers and noise	Is sensitive to outliers and noise but less sensitive than SL		

2.3 Consensus clustering

Different hierarchical clustering algorithms are proper for different shaped clusters, so each algorithm may produce different clusterings for a given data set. Thus, puts up the problem of choosing one of these clustering (which it is not a trivial task). Many contributions to this problem are addressed in Chapter 3, consisting in how to validate the clusterings using indices. Lately, many works have sought to combine different clusterings obtained by different algorithms and still get a best data clustering, designated by consensus clustering.

The idea is to capture the common structural aspects of the various clusterings producing a better clustering, which often means, a more stable, more robust and more consistent clustering.

Figure 2.1 illustrates a processing of the consensus clustering. First, it is applied a clustering algorithm, LAC, varying a parameter, h, to a data set, thus yielding different clusterings. From the partitions obtained by these clusterings, P_1, P_2, …, P_m, and from a Consensus Function is obtained the ensemble result or the consensus clustering.

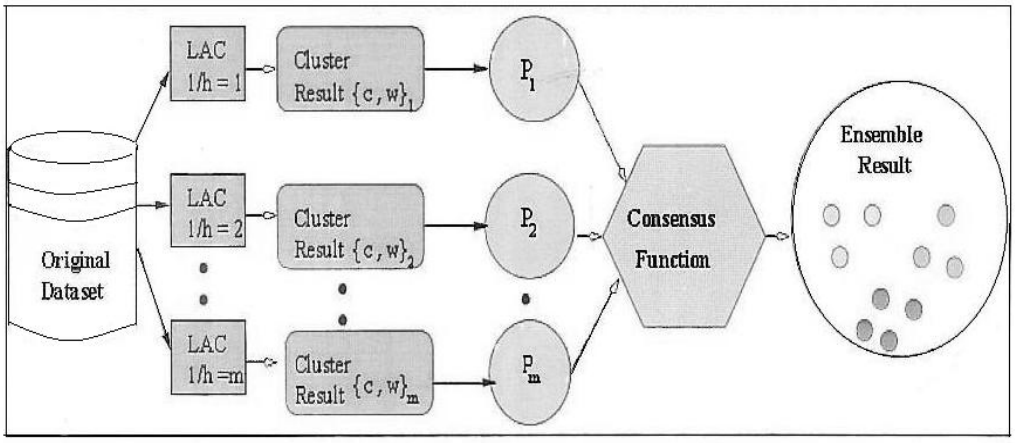

Figure 2.1- An illustrative figure of the consensus processing [2].

The various techniques in processing consensus clustering consist of two principal steps: Generation, which defines how to produce the set of individual clusterings or base clusterings to combine, and Consensus Function, describing how to combine the individual clusterings, finding the consensus clustering. Thus, different ways to obtain and combine clusterings lead to different consensus clustering techniques. Furthermore, each one of these techniques consider that certain properties (or objectives) should be fulfilled by the consensus clustering. Some of these properties are, 1) Stability- Lower sensibility to noise or outliers, 2) Consistency- A clustering similar to all the individual clusterings, 3) Robustness- Better performance than the individual clusterings and 4) Novelty- A clustering different from the individuals [97].

In the Generation step, there are no constrains about how the individual clusterings must be obtained. Therefore, different clustering algorithms or the same algorithm with different parameters/ initialization can be applied. A common idea among the different techniques is that the several clustering to combine must have a certain diversity between them, so that they provide more information in the processing of consensus [39]. On the second step, the Consensus Function focuses the methodology by combining these individual clusterings to obtain the consensus clustering. The Consensus Function is the main step for any consensus clustering technique and can be based, for instance, on Voting, Co-association Matrix, Graph and Hyper graph Partitioning, Information Theory, Finite Mixture Models and Genetic Algorithms. Moreover, some Consensus Functions are based on more than one of these approaches [97].

Next, we present some methodologies to obtain the consensus clustering, considered in literature as the traditional consensus clustering techniques.

Among several important contributions in the consensus clustering framework, it can be highlighted the works in [31], [33] and [87], [88]. These are the pioneers on traditional consensus clustering approaches and are perhaps, the most referred in literature. By such, we chose these consensus clustering techniques for our studies.

In [31], the consensus function is based on Voting and Co-association Matrix. The objective is to find consistent and robust consensus clustering. The individual clusterings are delivered by using the K-means algorithm. With the data clusterings obtained, pairs of elements are voted to be in the same cluster on consensus clustering when they belong to the same cluster in the different clusterings. The number of times that pair of elements are in the same cluster is counted and set on a matrix, the co-association matrix. This matrix can be viewed as a similarity measure between elements, and the consensus clustering is achieved by joining in the same cluster, pair of elements with a co-association value greater than 0.5 (the threshold pre-defined). This means that pairs of elements are in the same cluster in more than 50% of the individual clusterings.

As continuation of the work based on the voting mechanism, in [33] arises the concept of accumulation by evidence, EAC (Evidence Accumulation Clustering). It consists of a modification of [31] where the co-association matrix is represented as a graph. The idea is to cut weak links between nodes on graph, by a threshold called "highest lifetime", which corresponds to the minimum weight in the edges. This is analogous to cut the dendrogram produced by SL algorithm, being lifetime the range of threshold obtained by the distance between two consecutive levels on the dendrogram. One range with the highest value is selected, delivering the consensus clustering.

In order to build robust consensus clustering, in [87], [88], the authors propose a technique where the consensus clustering is achieved by an optimization problem, consisting on the consensus function maximization. The process is carried by applying Mutual Information and hyper graphs representation. The Mutual Information, concept from Information Theory [14] used to measure the shared information between pairs of clustering, is computed by the entropies. The consensus clustering is a clustering that shares most information with all possible clusterings. The objective of find clusters that maximize the Mutual Information by an exhaustive search of pairs of clusterings, raises computational problems. To solve this problem, three algorithms based on hyper graph representation and partitioning algorithms are proposed; CSPA - Cluster-based Similarity Partitioning Algorithm; HGPA – Hyper Graph Partitioning Algorithm and MCLA - Meta-Clustering Algorithm. These algorithms start from representing the clusterings in a hyper graph, where each clustering is represented by a hyper edge. In CSPA algorithm, first is constructed a co-association matrix. The entries of this matrix are weights associated to each two objects, corresponding on hyper graph representation, to the edge between the objects and the objects are the nodes. After it is applied, the graph partitioning algorithm, METIS [52]. This algorithm reduces the size of the hyper graph by collapsing the nodes and edges. With the reduced graph is applied a clustering algorithm obtaining a partition of the objects. METIS then extend the graph to construct a partition of the original graph leading to the consensus clustering. The greater the weight of the edge, greater is the similarity between objects. Thus, on the first phase of METIS, this is the criterion used to join the nodes having in common, edge with the highest weight. The partition obtained at the smaller graph, is by a clustering algorithm based on similarities. HGPA algorithm also applies a partitioning algorithm, HMETIS [51], but for partitioning hyper graphs. Eliminating the

minimal number of hyper edges (all hyper edges have the same weight) that corresponds to the relationships that occur less often. In MCLA algorithm is constructed a similarity matrix between clusters in terms of the amount of objects grouped in the respective clusters. In hyper graph representation the clusters are nodes and the edges between two nodes have weight which represents the similarity between the clusters. By the partitioning algorithm METIS, one obtains clusters called meta-clusters, and it is calculated the times that each object appears in a meta-cluster. Being each object assigned to the meta-cluster to which appears more often. Now, from these clusterings (associated to the three algorithms) is possible to search for final consensus clustering, the clustering which maximizes the Normalized Mutual Information. These authors, unlike those previous, use different algorithms to obtain the individual clusterings, and also pre define the desired number of clusters in the consensus clustering.

Further contributions to processing the consensus clustering have emerged in literature, having different commitments to obtain the consensus clustering. These commitments are regarding: the algorithms to obtain the individual clusterings; the way to represent these clusterings; the consensus function and the objectives to fulfill by the consensus clustering. The objectives most sought are that the consensus must be robust, consistent, and stable. Other objectives are considered such as, obtaining the consensus clustering with small cost or the consensus clustering must be a new clustering (different from the individual clusterings). Those contributions are summarized in Table 2.2.

Table 2.2: Comparison of some consensus clustering techniques referenced in literature.

References	Objectives	Individual clusterings	Clustering representation	Consensus function
Dimitriadou et al. (2001) [18]	Stability	Various algorithms	Set with all the clusters	Voting
Fred (2001) [31]	Consistence	K-means with different initializations	Co-association matrix	Voting
Strehl and Ghosh (2002) [87], [88]	Robustness and stability	Various algorithms and the same algorithm with different data (resampling)	Hyper graph	Hyper graph partitioning and Mutual Information
Topchy at al. (2003) [92]	Good performance, small cost and novelty	Weak clustering algorithms	New set characterizing the objects	Based on generalized Mutual Information
Ayad and Kamel (2003) [6]	-	K-means and graph partitioning	Hyper graph and similarity matrix	Shared nearest neighbour model
Topchy et al. (2004) [93]	Robustness, novelty and stability	K-means with different initializations	New set characterizing the objects	Statistics: maximum likelihood
Law et al. (2004) [61]	Novelty and robustness	Various algorithms and different initializations	Set with all the clusters	More stable clusters
Fern and Brodley (2004) [28]	Robustness	K-means with different data (resampling)	Hyper graph	Hyper graph partitioning
Fred and Jain (2005) [33]	Consistence and robustness	K-means with different initializations	Co-association matrix	Voting, SL and AL
Razgan and Domeniconi (2006) [2]	Robustness and stability	LAC algorithm with different data (resampling)	Hyper graph and set with all the clusters with weights	Hyper graph partitioning and more associated clusters
Faceli (2007) [26]	Robustness and stability	Various algorithms and different initializations	As population in genetic algorithm	Hyper graph partitioning and selection criterion of the genetic algorithms
Domeniconi and Razgan (2009) [20]	-	Various algorithms as hierarchical and K-means	Similarity matrix and hyper graph	Similarity and hyper graph partitioning

2.4 Conclusions

In this Chapter, we addressed the hierarchical clustering and the consensus clustering algorithms. These algorithms will be applied in our experiments in the next Chapters.

Chapter 3

Variability of hierarchical clustering algorithms

3.1 Summary

This Chapter is addressed to the subject validation of clusterings when several researches to validate the resulting clusterings analyse them in terms of stability or variability. We proceed to analyse the variability of the hierarchical clustering algorithms, referred in Chapter 2, exploring the profile of these clusterings. These clusterings are the base clusterings for the consensus clustering techniques application which are performed in Chapter 4.

3.2 Validating clusterings

Using different clustering algorithms for the same data set, or using the same clustering algorithm but with different initializations (or different parameters), can produce different clusterings. So, several studies have been concerned with validate the resulting clustering analysing them in terms of stability / variability.

The difficult task of choose one clustering can be based on evaluating the clustering's quality. The analysis of compactness and separation of the clusters does not always find the real clusters of a data set [11]. Furthermore, properties as variability or stability enable us to meet more stable solutions and infer about clustering quality.

Many works analyse the stability / variability / diversity of the clusterings obtained by data resampling for the purpose of validate clusterings. These works differ on the following issues:

i) The methodology for resampling data, as bootstrap in [53], [60] or cross-validation in [11], [55], [59], [75], [78];
ii) Clustering algorithm applied to the samples, as K-means and hierarchical algorithms in [55]; K-means and EM (Expectation Maximization) in [11]; K-means, EM and hierarchical algorithms in [60], [75]; or K-means, KNN (K-Nearest Neighbours) and hierarchical algorithms in [62];
iii) Validation indices, as Gap in [59]; Adjusted Rand in [11], [41], [55] or based on Information Theory in [11], [75];
iv) Validation criteria, as internal in [53], [55]; external in [41] or relative in [11].

Clustering validation can provide a quantitative answer through validation indices, for the need to validate the output of a clustering algorithm. Thus, a validity index can be seen as a factor which assesses the quality of a clustering [60].

The several approaches of clustering validation are based on indices or statistical measures, in accordance with the strategy adopted. Strategies or criteria can be classified in internal, external or relative.

Validation techniques that apply internal criteria, evaluate a clustering based on the data set as by the similarities matrix of the data, by the separability and homogeneity of the clusters. At this criteria, are applied indices such as, Gap [91] and Clest [35].

Techniques with external criteria, evaluate a clustering obtained, by the knowledge of the "real" clustering. Usually, the validity indices are based on the similarity measure between clusterings, as the indices Adjusted Rand [42], Normalized Mutual Information [87], [88], Jaccard [46], Folkes and Mallows [46], Hubert [46] and Dom [11].

On relative criteria, two clusterings obtained are compared, many times applying the same indices as in external criteria.

The Adjusted Rand index (ARI) and Normalized Mutual Information (NMI) are, perhaps, the most popular measures of similarity of clusterings.

The Rand index (1971) [77], measures the association between two clusterings and is calculated considering: i) Pairs of elements that are in the same cluster in a clustering and in the same cluster in the other clustering; ii) Pairs of elements that are in different clusters in a clustering and in different clusters in the other clustering. The Rand index had some problems, so, to solve them, in 1985 Hubert and Arabie [42] proposed the Normalized or Adjusted Rand Index (ARI).

Based on agreements and disagreements of two clusterings, to set the ARI, we consider a data set of *n* elements, and two different clusterings of the data, U and V. The clustering U with R clusters, $u_1, \dots, u_R$ and the clustering V with C clusters, $v_1, \dots, v_C$. The ARI value of these clusterings, can be obtained by the Equation 3.1, where the terms in the expression are: n_{ij} is the number of elements that are in cluster

u_i of the clustering U and in cluster v_j of the clustering V; $n_{i.}$ is the total of elements in cluster u_i and $n_{.j}$ is the total of elements in cluster v_j.

$$\mathrm{ARI(U,V)} = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \left[\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_{i=1}^{R}\binom{n_{i.}}{2} + \sum_{j=1}^{C}\binom{n_{.j}}{2}\right] - \left[\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}\right]/\binom{n}{2}} \tag{3.1}$$

ARI can take values since close to 0 (even negative values) until 1. The ARI value equals to 1 indicates perfect agreement between the clusterings, unlike values very close to 0 indicates total disagreement between the clusterings.

In Information Theory, the Normalized Mutual Information (NMI) is a symmetric measure to quantify the statistical information shared between two distributions [87], [88].

Considering the two clusterings U and V and the same descriptions of the terms in the ARI equation, NMI can be defined by the Equation 3.2.

$$\mathrm{NMI(U,V)} = \frac{-2\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{n_{ij}}{n}\ log\left(\frac{n_{ij}\,n}{n_{i.}n_{.j}}\right)}{\sum_{i=1}^{R}n_{i.}\ log\left(\frac{n_{i.}}{n}\right) + \sum_{j=1}^{C}n_{.j}log\left(\frac{n_{.j}}{n}\right)} \tag{3.2}$$

NMI can take values in the interval [0, 1]. The greater, the better, 1, indicates perfect agreement, otherwise, value 0 indicates that clusterings are totally independents from each other.

As our interest is on the variability of clusterings, we can mention some works concerning this that exist in literature. For instance, the works in [60], in which, the authors interpret a clustering algorithm as a statistical estimator and examine the variability of this estimator. This variability can be described as follows.

Consider a data set, Y, with size n. By resampling are obtained k sets of data samples, $Y^1, \dots, Y^k$, each one with the same size n. To each set of data sample is applied a clustering algorithm, designated by A, thus, obtaining, k clusterings, $A(Y^1), \dots, A(Y^k)$. The variability V of the clustering algorithm A is obtained by the Equation 3.3, where d measures the distance between clusterings and can be done by any measure of similarity between clusterings, as the indices Rand, Jaccard, Folkes & Malows and Hubert. Low values of V, mean small variability and hence that the clustering algorithm is stable.

$$\mathrm{V} = \frac{1}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} d(A(Y^i), A(Y^j)) \tag{3.3}$$

Another contribution to this issue is in [11]. These authors analyse the variability of a clustering by data resampling based on a weighted cross-validation procedure. They consider 20 weighted data samples and the original data sample. For each of them, they apply the clustering algorithm K-means to obtain the clusterings. After that, they calculate the agreement between the clustering of the original data sample and each one of the clusterings of the weighted data samples using the Adjusted Rand index. Once having the 20 values of the Adjusted Rand index, the standard deviation of them is used to measure the variability of the clustering.

3.3 Experimental design

In this section, we intend to analyse the variability of clusterings delivered by the traditional hierarchical algorithms. For that, we consider data resampling, and for each set of data sample we apply a hierarchical clustering algorithm to obtain the clusterings. Hence is calculated the agreement between them by the Adjusted Rand index (ARI) and relative criterion. The standard deviation of the ARI values measures the variability of the clustering, as in [11].

Also, we apply statistical analysis by hypothesis tests. The hypothesis under study is whether the different processing forms of the hierarchical clusterings or the different hierarchical clustering algorithms affect the respective variability.

To test this hypothesis, we conduct a set of experiments, which we start to describe.

3.3.1 Data sets

In order to reach the variety of situations regarding the data sets, we consider different simulated and real data sets. The differences between the data sets are related to cardinality, number of cluster, shape of the clusters and other characteristics such as close or well separated clusters and clusters with distinct densities. We also consider data sets with added noise and a data set with overlapping clusters. A description of each data set is given below.

Simulated data sets

In Figures 3.1 - 3.7 are represented the 2-dimensional simulated data sets used in our experiments and in Table 3.1 are the details of these data sets.

The data sets are, with random data, according to their partition into clusters, and Normal distribution. Some of them are data sets used by other authors. On some data sets, we introduce noise randomly uniformly distributed. There are seven data sets assigned, D1-4g, D2-3g, D2-3gr10 (data set D2-3g, with 10% noise), D3-3g, D3-3gr10 (data set D3-3g, with 10% noise), D4-10g [37] and D4-10gSS [37] (data set D4-10g, without overlapping clusters).

Real data sets

In our experiments we consider seven real data sets which were taken from the UCI Machine Learning Repository [94]. These data sets, besides different cardinalities, number of clusters and shape of the clusters, also have different dimensionality, in which, some of them are used in medical studies. These data sets are described below and summarized in Table 3.2.

- Iris: Refers to types of the iris flowers. The attributes are four: sepals length, sepals width, petals length and petals width. The clusters of iris are classified by, Setosa, Versicolour and Virginica.
- Ecoli: The clusters describe protein localization sites in Gram-negative bacteria E.coli [71].
- Wine: Consists of chemical analysis of thirteen constituents found on wines growing in the same region. The data clusters are according to the origin of the wine which can be from three different cultivars.
- Haberman's Survival: Contains cases from a study conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer. The attributes at time of the operation are: age of patient, year of the operation and number of positive

auxiliary nodes detected. The clusters are two, according to the patients' survival time, in which, one cluster has the patients that survived at least 5 years and the other cluster has the patients that do not survived 5 years.

- Blood: Taken from the Blood Transfusion Service Center in Hsin-Chu City in Taiwan. Were selected 748 donors at random from the donor database. The four attributes are: Recency – months since last donation, Frequency - total number of donation, Monetary - total blood donated, and Time – number of months since first donation. The data set is then divided into two clusters representing whether the donor donated blood in March 2007 (yes or no) [43].
- WDBC- Wisconsin Diagnostic Breast Cancer, contains 30 variables computed from digitized images of aspirated fine needle of a breast mass, describing the characteristics of a cell nuclei presents. The clusters are two, meaning that the diagnosis is benign or malignant [67].
- Breast Tissue: Consists of measures of electrical impedance of tissue samples taken freshly from the breast. This data set can be split into six clusters, Carcinoma, Fibro-adenoma, Mastopathy, Glandular, Connective and Adipose [81].

Table 3.1: Details of simulated data sets. Data generated by Normal distribution, $N(\mu, \sigma^2)$ where μ is the mean and σ^2 is the variance. D is the dimensionality, C is the number of clusters, Ni is the number of data of the cluster i, OC and AN means overlapping clusters and add data noise, respectively. The data noise are generated by Uniform distribution U(a,b) where (a,b) is the support interval.

Data set	D	C	Ni	Source	OC	AN
D1-4g	2	4	15×35 35×35	C1: $\mu_x = 0.5, \mu_y = 0, \sigma^2{}_x = \sigma^2{}_y = 0.05$ C2: $\mu_x = -1, \mu_y = 4, \sigma^2{}_x = \sigma^2{}_y = 0.2$ C3: $\mu_x = 2, \mu_y = 0, \sigma^2{}_x = \sigma^2{}_y = 0.2$ C4: $\mu_x = 2, \mu_y = 3.5, \sigma^2{}_x = \sigma^2{}_y = 0.2$	No	No
D2-3g	2	3	3×50	C1: $\mu_x = -1, \mu_y = 0, \sigma^2{}_x = \sigma^2{}_y = 0.25$ C2: $\mu_x = 1.5, \mu_y = 2.5, \sigma^2{}_x = \sigma^2{}_y = 0.25$ C3: $\mu_x = 8.5, \mu_y = 10, \sigma^2{}_x = 1.5, \sigma^2{}_y = 2.25$	No	No
D2-3gr10	2	3	50×56×59	C1: $\mu_x = -1, \mu_y = 00, \sigma^2{}_x = \sigma^2{}_y$ =0.25 C2: $\mu_x = 1.5, \mu_y = 2.5, \sigma^2{}_x = \sigma^2{}_y = 0.25$, U(3,4) C3: $\mu_x = 8.5, \mu_y = 10, \sigma^2{}_x = 1.5, \sigma^2{}_y = 2.25$, U(6,7)	No	Yes
D3-3g	2	3	3×100	C1: $\mu_x = -1, \mu_y = -1, \sigma^2{}_x = \sigma^2{}_y = 0.5$ C2: $\mu_x = 2, \mu_y = 2, \sigma^2{}_x = \sigma^2{}_y = 0.7$ C3: $\mu_x = -3, \mu_y = 3, \sigma^2{}_x = \sigma^2{}_y = 0.1$	No	No
D3-3gr10	2	3	130 100×100	C1: $\mu_x = -1, \mu_y = -1, \sigma^2{}_x = \sigma^2{}_y = 0.5$, U(0,0.3) C2: $\mu_x = 2, \mu_y = 2, \sigma^2{}_x = \sigma^2{}_y = 0.7$ C3: $\mu_x = -3, \mu_y = 3, \sigma^2{}_x = \sigma^2{}_y = 0.1$	No	Yes
D4-10g	2	10	25×5 50×5	Ci: $\mu_x, \mu_y \in [0, 50]$; $\sigma^2{}_x = \sigma^2{}_y \in [0.1, 0.3]$, i=1,..10	Yes	No
D4-10gSS	2	10	25×5 50×5	Ci: $\mu_x, \mu_y \in [0, 50]$; $\sigma^2{}_x = \sigma^2{}_y \in [0.1, 0.3]$, i=1,..10. For each 2 clusters, $d(C_k, C_l) > 3(\sigma_k + \sigma_l)$ where C_k and C_l are the centre points and σ_k and σ_l are the standard deviations, respectively	No	No

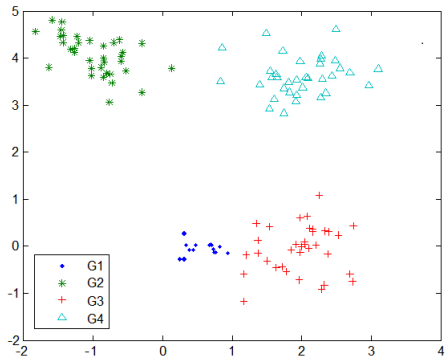

Figure 3.1- Representation of the data set D1-4g.

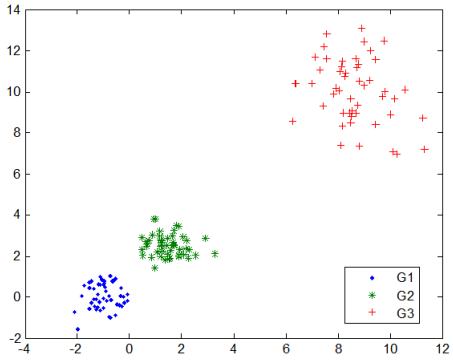

Figure 3.2- Representation of the data set D2-3g.

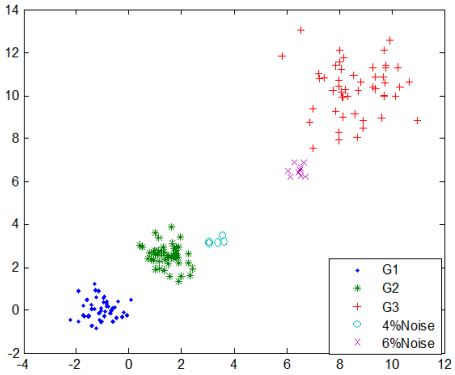

Figure 3.3- Representation of the data set D2-3gr10.

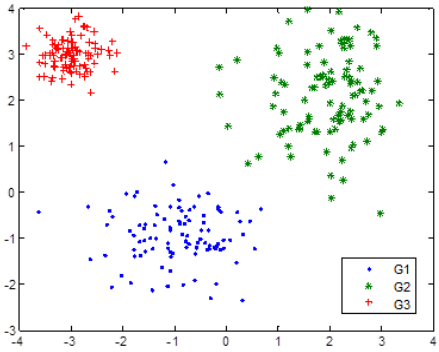

Figure 3.4- Representation of the data set D3-3g.

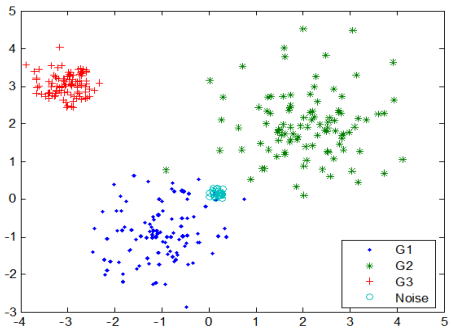

Figure 3.5- Representation of the data set D3-3gr10.

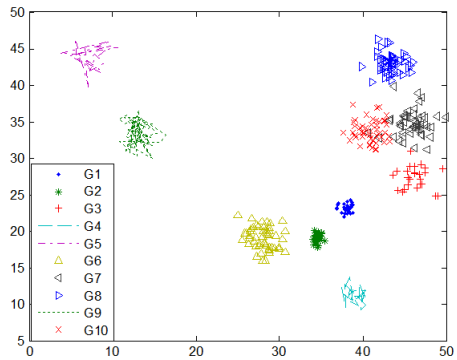

Figure 3.6- Representation of the data set D4-10g.

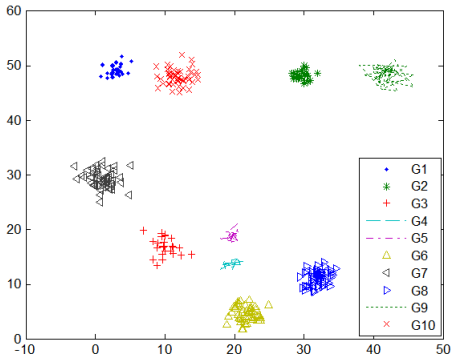

Figure 3.7- Representation of the data set D4-10gSS.

Table 3.2: Summary of the real data sets. N is the cardinality of data set, C is the number of clusters and D is the dimensionality.

Data set	N	C	D
Iris	150	3	4
Ecoli	336	8	7
Wine	178	3	13
Haberman's Survival	306	2	3
Blood	748	2	4
WDBC	569	2	30
Breast Tissue	106	6	9

3.3.2 Generation of the hierarchical clusterings

To obtain the clusterings, we apply, SL, CL, AL and W, the hierarchical clustering algorithms (with the Euclidean distance) to each sample of the data sets.

For each data set, we consider data resampling without replacement, yielding 50 sets of data samples, each one with size (2⁄3)N, where N is the cardinality of the data set. For the real data sets, before the resampling, the data are normalized having mean 0 and standard deviation 1. Each hierarchical clustering algorithm is applied to these data samples, obtaining the corresponding set of 50 clusterings.

As hierarchical clustering algorithms produce a hierarchy of partitions, the clusterings are obtained by cutting the hierarchy at a level in accordance with the number of clusters of the known data structure.

3.4 Variability analysis

Once obtained 50 clusterings for each data set and each hierarchical clustering algorithm, we calculate the agreement between these clusterings in pairs by the ARI. So, once having the 1225 values of the ARI, we calculate the average, as well as the standard deviation of them. The variability measure of the hierarchical clustering algorithms is described below and the results for each data set and each hierarchical clustering algorithm are stated in Table 3.3.

Considering a data set *Y* with size N, by resampling is obtained 50 sets of data samples, $Y^1, \dots, Y^{50}$, each one with the same size 2N/3. To each set of data samples is applied a hierarchical clustering algorithm, in which we designate by A (SL, CL, AL and W), thus obtaining 50 clusterings, $\mathrm{A}(Y^1), \dots, \mathrm{A}(Y^{50})$. The variability, V, of the clustering algorithm A is obtained by the Equation 3.4 and the ARI by the Equation 3.1.

$$\mathrm{V} = \sqrt{\frac{1}{1224}\sum_{i=1}^{49}\sum_{j=i+1}^{50}\left(\mathrm{ARI}(\mathrm{A}(Y^i),\mathrm{A}(Y^j)) - \frac{1}{1225}\sum_{i=1}^{49}\sum_{j=i+1}^{50}\mathrm{ARI}(\mathrm{A}(Y^i),\mathrm{A}(Y^j))\right)^2} \tag{3.4}$$

Also is considered the differences of the variabilities of the hierarchical clustering algorithms applying statistical inference. Assuming the normality of the data, for each data set, we apply the hypothesis test (unilateral) of equality of variances, the F Snedecor statistic, considering the significance level set to 5%. Thereby, we can statistically conclude about the relation of the variances of the different hierarchical clustering algorithms. On Table 3.4 are displayed these relations.

Table 3.3: Comparison between the hierarchical clustering algorithms in terms of the ARI average and the variability, for each data set. The best relative results are highlighted.

	Data set	Algorithm	Average	Variability
Simulated data sets	D1-4g	SL	0.9119	0.0928
		CL	0.9672	0.0583
		AL	0.9950	0.0185
		W	0.9857	0.0438
	D2-3g	SL	0.8098	0.2247
		CL	0.9437	0.0399
		AL	0.7024	0.2113
		W	1	0
	D2-3gr10	SL	0.9104	0.1081
		CL	0.7056	0.2526
		AL	0.8570	0.1972
		W	0.9983	0.0085
	D3-3g	SL	0.7631	0.2121
		CL	0.9596	0.0440
		AL	0.9852	0.0262
		W	0.9875	0.0190
	D3-3gr10	SL	0.9108	0.1560
		CL	0.8240	0.1488
		AL	0.9855	0.0291
		W	0.9657	0.0722
	D4-10g	SL	0.9652	0.0554
		CL	0.9127	0.0603
		AL	0.9279	0.0532
		W	0.9532	0.0323
	D4-10gSS	SL	0.9881	0.0250
		CL	0.9927	0.0104
		AL	0.9971	0.0052
		W	0.9952	0.0080
Real data sets	Iris	SL	0.9683	0.0409
		CL	0.5345	0.2241
		AL	0.9276	0.1045
		W	0.7637	0.1985
	Ecoli	SL	0.8675	0.0857
		CL	0.5934	0.1397
		AL	0.8477	0.0787
		W	0.5864	0.1164
	Wine	SL	0.5893	0.3922
		CL	0.4108	0.1834
		AL	0.4648	0.3834
		W	0.8202	0.0826
	Haberman's Survival	SL	0.5570	0.4780
		CL	0.6326	0.3401
		AL	0.6522	0.3638
		W	0.3055	0.3293
	Blood	SL	0.8163	0.3912
		CL	0.7965	0.3188
		AL	0.8062	0.3770
		W	0.4657	0.2391

	WDBC	SL	0.5304	0.5045
		CL	0.5258	0.4693
		AL	0.6125	0.4625
		W	0.6361	0.1392
	Breast Tissue	SL	0.6924	0.2655
		CL	0.6862	0.1720
		AL	0.8230	0.1626
		W	0.6692	0.1714

Table 3.4: Relations between the variances of hierarchical clustering algorithms by the F Snedecor statistical test, for each data set.

Data set	Relations
D1-4g	SL>CL>W>AL
D2-3g	SL=AL>CL>W
D2-3gr10	CL>AL>SL>W
D3-3g	SL>CL>AL>W
D3-3gr10	SL=CL>W>AL
D4-10g	SL=CL=AL>W
D4-10gSS	SL>CL>W>AL
Iris	CL=W>AL>SL
Ecoli	CL=W>SL=AL
Wine	SL=AL>CL>W
Haberman's Survival	SL>CL=AL=W
Blood	SL=CL=AL>W
WDBC	SL=CL=AL>W
Breast Tissue	SL>CL=AL=W

Analysing the variability results in Tables 3.3 and 3.4, at almost all the cases, the clustering algorithms presenting greater average of ARI values also presents the lowest variability, with exceptions for the simulated data set D4-10g, and the real data set Blood.

Our interest is to compare the variability of the hierarchical clustering algorithms for each data set. Henceforth, referring to the variability of a clustering algorithm as greater or lower, we want to say that is in relation to the other clustering algorithms.

Considering the simulated data sets, W or AL algorithms, feature the lower variabilities for all the cases. For one data set, W presents 0 variability and ARI average equals to 1. From seven data sets, looking for the lower variability, we find that four of them are from W algorithm, namely D2-3g, D2-3gr10, D3-3g and D4-10g, other three from AL algorithm, as D1-4g, D3-3gr10 and D4-10gSS. By other hand, SL presents the greater variability for almost all the data sets (with exception of D2-3gr10 data sets).

For almost all the simulated data sets, the different algorithms feature very different variabilities between them, excluding the data sets, D1-4g, D4-10g and D4-10gSS.

Analysing the effect of data noise (D2-3gr10 and D3-3gr10) on the variability, CL algorithm shows the biggest relative sensitivity to the noise. And all the algorithms are affected by the existence of overlapping clusters (D4-10g).

Concerning to the real data sets, W algorithm presents, at almost all the cases (five data sets), the lower variability. For some data sets, more than one algorithm has smaller variability. For instance, for the data set Ecoli, SL and AL feature equally the smaller variability. Also for data sets Haberman's Survival and Breast Tissue, CL, AL and W feature equally the smaller variability.

In general the algorithms which present higher variabilities, present lower average ARI values for the simulated and the real data sets.

For some data sets all the algorithms present very high variabilities, as for Haberman's Survival and Blood data sets.

Also we compare the average ARI values by the statistical hypothesis test for difference of means. For this test the statistic applied has asymptotic Normal distribution and is considered the significance level set to 5%. We find that W algorithm presents relatively higher average of the ARI value (bigger than 0.95) for all the simulated data sets, but does not for the real data sets.

By the experimental results we can state that, for each data set, the hierarchical clustering algorithms have different variability between them.

Now, analysing the graphic representation (see Figures 3.1 - 3.7), as the characteristics of the simulated data sets (see Table 3.1), and regarding the properties of the hierarchical clustering algorithms, as well as the result of their variability, we can establish the following:

- Considering the data set D1-4g, wherein 3/4 of the clusters (C2, C3 and C4) despite having the same cardinality and cohesion, furthermore, they have greater variance than the remainder cluster. So, they are neither compact nor elongated. It's expected that SL and CL produce less stability, and is mainly due to the result of its higher variability in relation to the other algorithms.
- For data set D2-3g, having all clusters the same cardinalities, 2/3 of them (C1 and C2) have smaller variance than the remaining one, they are then more compact, also small with spherical shape and close to each other. After this, is expected that CL and W present more stability, according to the lower variability of these algorithms in relation to SL and AL.
- With regard to data set D3-3g, where all the clusters have the same cardinality and spherical shapes, 2/3 of them (C1 and C2) are less compact than the remaining one, also slightly apart and having larger diameters. It is expected that SL are less stable and moreover, present a higher variability's value compared to the other algorithms.
- Regarding the data set D4-10gSS (without overlapping clusters), wherein the different clusters have different cardinalities, in general, they are compacts and some of them slightly separated. It is expected that SL presents less stability, resulting in higher variability value with regard to the remaining algorithms.

- As CL is more sensitive to outliers or data noisy, the variability values for data sets D2-3gr10 and D3-3gr10, are expected, presenting the highest variability among the other algorithms.

Faced with the results delivered, we can confirm the hypothesis under consideration, that different processing of the hierarchical clustering can influence the respective variability.

3.5 Conclusions

In this Chapter, we proposed to analyse empirically the variability of the hierarchical clustering algorithms, such as, Single Linkage, Complete Linkage, Average Linkage and Ward.

The variability of the clustering algorithms is measured by the Adjust Rand index, more precisely by the standard deviation of the ARI values. The clusterings were obtained by those algorithms applied to data resampling of synthetic and real data sets.

This study was performed to verify a hypothesis test about the difference of variability on the hierarchical algorithms. The analysis of the known properties of the hierarchical clustering algorithms leads to the identification of a new property of these algorithms based on the variability.

Applying a hierarchical algorithm better suited to a data set with certain characteristics regarding to its clusters, this algorithm presents less variability. As for instance, SL favours connectivity, arbitrary shape, elongated and well separated clusters, in the same circumstances, SL presents lower variability. CL favours compactness, spherical shape, small and close clusters and also in this circumstance, CL presents lower variability.

Through these researches we searched to define profiles of clusterings in terms of their variability, in which these clusterings will be the base clusterings for the consensus. The application of consensus clustering techniques to these base clusterings are performed in Chapter 4.

Chapter 4

Validation of consensus clustering

4.1 Summary

In this Chapter we address the subject validation of consensus clustering, as well as some works intend to achieve the best consensus clustering. We analyse the performance of the traditional consensus clustering techniques applied to some sets of base clusterings. Whereas each set of base clusterings has a known profile in terms of variability of their clusterings. The studies concerning on clusterings' variability were performed in Chapter 3.

4.2 Related works

Faced with the existence of different consensus clustering techniques, some works have been concerned about validating the resulting consensus clustering. We describe below some proposed researches comparing the performance of the different consensus clustering. These comparisons are taking into account some measures for the purpose of identifying the individual clusterings (or base clusterings) that leads to the best consensus clustering.

Considering,

- Y a data set of n elements with some data structure into clusters;
- $P_i = \{C_{1_i}, \dots, C_{K_i}\}$ a clustering of Y into K_i clusters;
- $P = \{P_1, \dots, P_m\}$, a set of base clusterings with m clusterings of Y;
- P^* a consensus clustering;
- P^T the true clustering of the data set.

In [39], the authors propose four diversity measures for the base clusterings and the consensus clustering, based on the ARI. The base clusterings are obtained by K-means algorithms, with different initializations, and the consensus clustering is obtained by the EAC technique. The accuracy of a consensus clustering is with respect to a known true clustering of the data.

Formally, the first diversity measure is defined as the average diversity between each clustering, $P_i \in P$, and the consensus clustering, P^* (see Equation 4.1) where $AR(P_i, P^*)$ is the ARI value of the pairs of clusterings P_i and P^*.

The second measure is defined as the standard deviation of the first diversity (see Equation 4.2).

The third and forth diversity measures are derived from the first and second ones, and can be seen in Equations 4.3 and 4.4, respectively.

The accuracy of the consensus clustering, P^*, is calculated as, $AR(P^T, P^*)$.

$$Div_1(P, P^*) = \frac{1}{m}\sum_{i=1}^{m}\left(1 - AR(P_i, P^*)\right) \tag{4.1}$$

$$Div_2(P, P^*) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(1 - AR(P_i, P^*) - Div_1(P, P^*)\right)^2} \tag{4.2}$$

$$Div_3(P, P^*) = \frac{1}{2}\left(1 - Div_1(P, P^*) + Div_2(P, P^*)\right) \tag{4.3}$$

$$Div_4(P, P^*) = \frac{Div_2(P, P^*)}{Div_1(P, P^*)} \tag{4.4}$$

All these measures are compared and the authors conclude that only the first and the third measures present some relation with the consensus clustering quality. Moreover, they conclude that one should select the base clusterings with median value of the diversity to get the best consensus clustering.

The same authors in another work [38] evaluate the accuracy of the consensus clustering using 24 different scenarios, each one describing the base clusterings algorithms and the consensus function applied. The base clusterings algorithms used are: K-means, SL, AL applied to the data sets and also considering samples of the data sets. The consensus functions derive from the algorithms: CSPA, HGPA, by co-association matrix and by a matrix representing the data rather than similarities. The accuracy of the consensus clustering is obtained like in [39]. After performing a set of experiments according to the different scenarios, were taken some conclusions. These are: the best consensus clustering is achieved by using base clusterings obtained by K-means algorithms, and by the consensus function that interprets the consensus matrix of the base clusterings as data instead of similarity.

In [21] a new measure is proposed to select the best consensus clustering among a variety of them. It is based on the concept of Average Cluster Consistency, ACC, which provides the average similarity between each clustering, P_i, of base clusterings and the consensus clustering, P^*. This measure is defined by the Equations 4.5 and 4.6, where, $K_i \geq K^*$, being K_i and K^*, the number of clusters of the clustering P_i and P^*, respectively. $\left|Inters_{kj}\right|$ is the cardinality of the set of common data to the j^{th} and k^{th} clusters of the clustering, P_i and P^*, respectively.

The quality of the consensus clustering, P^*, is calculated by the Consistency index, Ci, which measures the quantity of data shared in matching clusters of the true clustering and the consensus clustering. This index is defined by the Equation 4.7, where K^T is the number of clusters of the true clustering and $\left|C_{K^*} \cap C_{K^T}\right|$ is the cardinality of P^* and P^T , K^{th} matching clusters data patterns intersection [21].

$$ACC(P, P^*) = \frac{1}{m}\sum_{i=1}^{m} sim(P_i, P^*) \tag{4.5}$$

$$sim(P_i, P^*) = \frac{1}{n}\sum_{j=1}^{K_i} \max_{1\le k\le K^*} \left|Inters_{kj}\right| \left(1 - \frac{|C_{K^*}|}{n}\right) \quad (4.6)$$

$$Ci(P^T, P^*) = \frac{1}{n} \sum_{k=1}^{min\{K^*, K^T\}} \left|C_{K^*} \cap C_{K^T}\right| \quad (4.7)$$

In the experiences, the base clusterings are obtained, among other algorithms, by K-means, SL, AL, CL, and also considering join clusterings obtained by these algorithms. The number of clusters is randomly chosen between 10 and 30. The consensus clustering is obtained by EAC technique and also by other two variants of WEACS technique (extension of EAC considering weights to the voting mechanism). The accuracy of the consensus clustering is measured by Ci (in Equation 4.7), with respect to a known true clustering of the data. The authors conclude that the best consensus clustering is the one that achieves the highest value of ACC (in Equation 4.5).

4.3 Experimental analysis

In this Section, according to the variability of the hierarchical clustering algorithms, we propose to analyse its implications on the performance of the consensus clustering techniques.

The performance of a consensus clustering technique is measured by the match between the consensus clustering obtained and the known true clustering. For this, we apply the Adjusted Rand index (ARI) and Normalized Mutual Information (NMI). While ARI quantifies the proportion of pairs in agreement of two clustering, NMI informs if two clustering are independent one from another.

For our studies, we apply hypothesis tests. The hypothesis under study is whether the performance of the consensus clustering techniques depends on the variability of base clusterings. To test this hypothesis, we perform a set of experiments, which are described as follows.

We consider the same data sets and clusterings reported in Chapter 3. Thus, for each data set and each hierarchical clustering algorithm, we have 50 clusterings, which are the base clusterings to obtain the consensus clustering.

We apply the traditional consensus clustering techniques, referred in Chapter 2. One based on Voting scheme [31] (TEC.1), other is based on co-association matrix, Evidence Accumulation Clustering [33] (TEC.2) and another is based on Mutual Information and hyper graphs [87], [88] (TEC.3).

4.4 Impact of base clusterings variability on consensus

In order to compare the consensus clusterings obtained from the techniques, is calculated the ARI and also the NMI between the consensus clustering and the known clustering of the data sets. For each data set and each set of base clusterings derived from a hierarchical clustering algorithm, the Table 4.1 contains the ARI and NMI values of each consensus clustering technique.

By observing the results in Table 4.1, one can establish the possible differences of the consensus techniques and still that, some technique features better performance than other techniques in conformity with their ARI and NMI values. Besides, these indices have very similar behavior.

For some data sets, as D3-3g and D4-10gSS, the TEC.3 outperforms the other techniques whatever it is the base clusterings algorithm. For some other data sets, in no situation some technique outperforms the others, as for instance, Haberman's

Survival, Blood and Breast Tissue data sets. Besides, for these data sets no technique presents good performance.

Table 4.1: Comparison between the performances of consensus clustering techniques. The best relative results are highlighted.

	Data set	Clustering	ARI			NMI		
			TEC.1	TEC.2	TEC.3	TEC.1	TEC.2	TEC.3
Simulated data sets	D1-4g	SL	0.5520	0.8265	0.9752	0.6756	0.8999	0.9716
		CL	0.7234	0.9823	0.9823	0.7678	0.9743	0.9743
		AL	0.7956	0.9823	0.9823	0.8215	0.9743	0.9743
		W	0.7164	0.9823	0.9823	0.7762	0.9743	0.9743
	D2-3g	SL	0.8310	0.5584	1	0.8165	0.7424	1
		CL	0.3090	0.5681	0.4934	0.4742	0.7612	0.5795
		AL	0.8500	0.5681	1	0.8327	0.7612	1
		W	0.7901	1	1	0.7865	1	1
	D2-3gr10	SL	0.2845	0.4183	0.4115	0.3935	0.4955	0.4806
		CL	0.4741	0.4183	0.7937	0.5760	0.4955	0.7873
		AL	0.2737	0.4183	0.3605	0.4076	0.4955	0.4134
		W	0.5904	0.7937	0.7937	0.6282	0.7873	0.7873
	D3-3g	SL	0.8521	0.5698	0.9801	0.8095	0.7612	0.9702
		CL	0.8477	0.5698	0.9801	0.8117	0.7612	0.9702
		AL	0.8813	0.5698	0.9801	0.8392	0.7612	0.9702
		W	0.8853	0.5698	0.9801	0.8448	0.7612	0.9702
	D3-3gr10	SL	0.5072	0.5438	0.6021	0.6064	0.7500	0.6581
		CL	0.6511	0.5438	0.9628	0.7273	0.7500	0.9516
		AL	0.8437	0.9628	0.9628	0.8027	0.9516	0.9516
		W	0.8241	0.5438	0.9628	0.7774	0.7500	0.9516
	D4-10g	SL	0.6781	0.7731	0.7604	0.8236	0.9279	0.8931
		CL	0.7186	0.7731	0.9247	0.8291	0.9279	0.9514
		AL	0.7612	0.9142	0.9518	0.8482	0.9712	0.9728
		W	0.7892	0.7731	0.9382	0.8529	0.9279	0.9594
	D4-10gSS	SL	0.8571	0.9142	0.9835	0.8816	0.9712	0.9845
		CL	0.8748	0.9142	0.9440	0.9017	0.9712	0.9551
		AL	0.8584	0.9142	1	0.8937	0.9712	1
		W	0.8531	0.9142	0.9875	0.8874	0.9712	0.9862

Real data sets	Iris	SL	0.4560	0.5584	0.5572	0.5786	0.7424	0.6999
		CL	0.3368	0.0004	0.5897	0.5119	0.4687	0.6226
		AL	0.4436	0.5681	0.5601	0.5616	0.7612	0.7187
		W	0.4712	0.5681	0.6440	0.5810	0.7612	0.6845
	Ecoli	SL	0.0440	0.0407	0.0171	0.2291	0.2278	0.0837
		CL	0.2943	0.0381	0.6579	0.5383	0.2105	0.6809
		AL	0.5706	0.0381	0.4761	0.6155	0.2105	0.6064
		W	0.1579	0.0381	0.5043	0.5247	0.2105	0.6226
	Wine	SL	-0.0142	-0.0083	-0.0078	0.0909	0.0645	0.0215
		CL	0.3691	0.0009	0.7497	0.5686	0.4560	0.7421
		AL	-0.0062	-0.0020	-0.0115	0.1423	0.0267	0.0684
		W	0.5716	0.4394	0.8185	0.6528	0.5865	0.8080
	Hab. Survival	SL	0.0332	0.0073	0.0072	0.0814	0.0336	0.0055
		CL	0.0581	0.0030	0.0947	0.0981	0.0006	0.0469
		AL	0.0132	0.0002	0.0368	0.0710	0.3138	0.0299
		W	0.0326	0.00003	0.0046	0.1372	0.3179	0.0063
	Blood	SL	-0.0137	-0.0036	-0.0036	0.0231	0.0072	0.0072
		CL	0.0272	0.0311	0.0311	0.0743	0.0350	0.0350
		AL	0.0096	0.0311	0.0311	0.0611	0.0350	0.0350
		W	0.0218	-0.00001	0.0293	0.0668	0.2861	0.0060
	WDBC	SL	0.0042	0.0048	0.0058	0.0603	0.0280	0.0126
		CL	0.0150	0.0048	0.0277	0.0650	0.0280	0.0773
		AL	0.0019	0.0048	0.0043	0.0575	0.0280	0.0051
		W	0.5696	-0.00001	0.6371	0.4397	0.3227	0.5120
	Breast Tissue	SL	0.0259	0.0007	0.0305	0.3014	0.1755	0.1613
		CL	0.2111	-0.0017	0.2610	0.5509	0.0487	0.4623
		AL	0.1214	0.1615	0.1768	0.4316	0.4538	0.3946
		W	0.1521	0.1671	0.2620	0.5261	0.4606	0.4980

Also observing the variability of the base clusterings derived from the different algorithms studied in Chapter 3 (Section 3.4), we can establish the following:

- Considering the simulated data sets, D1-4g, only for base clusterings obtained by SL there are differences at the three techniques. Actually, TEC.3 presents good performance and outperforms the other techniques and we note that, SL presents statistically greater variability than the remaining algorithms.
- Regarding the data set, D2-3g, whereas TEC.2 outperforms the other techniques with base clusterings obtained by CL. But, TEC.3 presents good performance outperforming the other techniques, considering SL or AL. On the other hand, SL and AL statistically have the same variability as also greater than the remaining algorithms and CL statistically has moderate variability.

- For D2-3gr10, TEC.2 outperforms the other techniques with base clusterings obtained by SL or AL. Also, TEC.3 presents good performance and outperforms the other ones, considering CL which statistically has greater variability than the remaining algorithms. SL and AL present moderate variability.
- As regard to D3-3gr10, TEC.3 presents good performance outperforming the other techniques with base clusterings obtained by SL or CL or W, which statistically have greater variability than AL.
- Considering the real data set Iris, TEC.2 outperforms the other techniques with base clusterings obtained by SL or AL. Besides, TEC.3 features better performance than the other techniques, with CL or W, which, statistically have equally variability as greater than the remaining algorithms.
- Observing the data set Ecoli, TEC.1 has the best performance, relatively to the other techniques, with AL. Moreover, TEC.3 outperforms the other ones with CL or W, which, have the same variability and greater than the remaining algorithms.
- For data set Wine, TEC.3 shows better performance than the other techniques, with CL or W which have lower variability relatively the remaining algorithms. While, for data set WDBC, TEC.3 shows better performance than the other techniques, with W having also lower variability relatively the remaining algorithms.

Thus, one can acknowledge that, TEC.3 of consensus clustering outperforms the other techniques, when it is applied to the base clusterings having greater variability relatively to the others, notably for the data sets, D1-4g, D2-3g, D2-3gr10, D3-3gr10, Iris, and Ecoli.

Also, TEC.2 prevails with algorithms having moderate variability, for the data sets D2-3g, D2-3gr10. For the data sets, as, D3-3g and D4-10gSS, TEC.3 outperforms the other techniques independently of the algorithms applied. About the data sets, Haberman's Survival, Breast Tissue and Blood, the three techniques show approximately the same performance assuming any of the algorithms.

Thereby, we can assert that, when there are differences on the performances of consensus clustering techniques, TEC.3 presents good performance and better

performance, relatively to other techniques, independently of the algorithms (this is observed in two data sets). Or, TEC.3 presents good performance and better performance, relatively to other techniques, with algorithms having greater variability relatively to the other algorithms. This happen in four out of seven simulated data sets and two out of the four real data sets. The data sets excluded of the statements above have a known data clustering with overlapping clusters (D4-10g) or have high dimensionality (WDBC).

So, for some data sets, we can confirm the hypothesis under consideration, in which, the performance of some consensus clustering techniques as TEC.3, depends on the variance of the base clusterings provided by a hierarchical algorithm. Thus, the consensus clustering provided by this technique can be evaluated by the knowledge of the variance of the base clusterings.

4.5 Conclusions

Several approaches to create consensus clustering are proposed and carried out in many ways which may lead to different consensus clustering for the same base clusterings. Thus, some works to evaluate/select the best consensus clustering have been proposed in literature, such as measures of the diversity [39], or consistency [21] between the base clusterings and the consensus clustering. These works evaluate the consensus by measures between the base clusterings and the consensus clusterings. For instance, regarding the works in [39], while the authors calculate the diversity measures between the base clusterings and the consensus clustering, in our analysis we calculate the variability measure between all the clusterings of the set of base clusterings. Moreover, one of the diversity measures is the standard deviation of the other diversity, based on $1 - AR(P_i, P^*)$. At our works, the variability of the base clusterings are calculated by the standard deviation of the ARI values, based on $AR(P_i, P_j)$.

Through our researches we intended to explore the profiles of base clusterings obtained by the hierarchical algorithms in function of their variabilities (in Chapter 3) and from these profiles, decide which consensus clustering technique to apply.

These studies were performed by experimentally verifying a hypothesis under consideration, which is the possibility of choosing the most appropriate consensus clustering, according to a particular type of variances of the base clusterings. Actually, when the consensus techniques present different performances, in most of the cases the technique based on Mutual Information and hyper graphs presents good performance and furthermore outperforms the others. This is achieved considering a set of base clusterings, where the clusterings are provided by a hierarchical clustering algorithm and having between them relatively higher variances. Thus, we found a condition which conducts to the existence of the consensus clustering, as well as, a studied strategy to evaluate the consensus clustering.

Chapter 5

Context of biometrics for recognition

5.1 Summary

In this Chapter, we present the real application that is used in Chapters 6 and 7, the hands biometrics for recognition.

In the contextualization of hands biometrics for recognition, we present: a literature review approaching the more cited works over the time, since the first systems created for recognition; a study about the number of published papers addressing this subject and finally some conclusions of our studies.

5.2 Introduction

The identification of people is a problem with considerable importance for economic and public safety. For example, if a person can take our identity, our bank balance can be diverted and our permission might be used to access certain places to commit terrorist acts.

The recognition of the individuals, based on the facial image or other characteristics, is a natural and simple process when performed by humans. Allowing quickly to identify any person and noticing her/his emotional state, even at the most diverse conditions, such as variations in light, distortion or deformation. Nevertheless, the problem is that it's expensive and doesn't allow the access to the databases.

Furthermore, the human visual recognition also has issues. An important one is the increase of failures in phenotypes, in relation to people with who we have less contact. Thus, it is known the greater difficulty in to identify people having different race from ours. This can cause problems, for example, in airports [72].

The identification using automatic/computer systems is therefore a field of research that surely is having huge impact on our society because it is potentially cheaper, allows quickly processing many people (crowds). Moreover it makes a correspondence between each person and a database, for example to make an automatic payment or to seek an outlaw by the police. Furthermore, it has the ability to analyse data, e.g. fingerprints, which the human eye isn't capable of doing.

However, the use that is made of the information collected automatically brings ethical problems [74] and distrust in people. So, as a rule, automatic systems can only proceed to the identification after the person (or the judicial system) authorizes.

As the human recognition is time consuming, costly and prone to failure, for several years it has been developed recognition technologies based on facial photographs. These technologies are used for: combating false passports; supporting the law enforcement; the identification of missing people and for the minimization of

identity fraud [12]. The use of these technologies is growing rapidly, especially since the "September 11".

Currently, the capacity of automatic recognition is measured by the percentage of correct identification of people (error type 1- says that it is the person, and it is not; error type 2- says it is not the person, and it is). These measures still do not allow its application on a large scale, but many advances have been achieved [8].

All the biometrics information can be used for automatic identification. First, from a photograph or video image of people, it is measured the facial image characteristics which have the discriminatory power (e.g., the ratio of the distance between the eyes and the height of the nose). Second, it is verified if this measure corresponds to the previously value obtained in which is in a database or in a microchip card (e.g., passport). Also can be used more invasive information, such as the fingerprint or iris image of the eye.

The biometrics continuously study new physical or behavioural characteristics of living beings, including people in order to be able to identify them uniquely [45]. In biometrics studies characteristics from various parts of the body are identified, such as the eyes, the palm, the fingers, the retina or the eye's iris shape and even, teething (which currently is used in the identification of rotting corpses).

Several researches have sought characteristics that, besides having discriminatory power, are also observable in an efficient, fast, low cost, and also be stable over time, such as the hands biometrics.

5.3 Hands biometrics

Researchers in the biometrics field discovered that, by the hands biometrics it's possible to identify people [30]. People's hands, differ in their size and shape, and these differences can be used to distinguish one individual from another. Besides, the hand recognition systems have little cost, needing only a camera or scanner, providing fast results and can achieve great percentage of recognition. Another important factor is that the recognition by the hand is only possible if the person authorize, thus doesn't bring ethical problems.

The general recognition systems based on hands biometrics consist on three main steps, firstly the obtainment of the hands images, secondly the features extracted and finally, the feature matching or recognition. Different hands biometrics recognition systems have different commitment regarding the steps above. In Figure 5.1 is a general representation of these systems.

For the hands image acquisition, initially the researchers used a digital camera and the hands were placed on a support with pins, conditioning the position of the palm and the fingers. Lately, the digital camera and the pins are replaced by a document scanner.

For the features extraction, different hands biometrics have been used, such as hand geometry, hand shape, palm print, hand vein, vascular patter of hand, finger print, finger knuckle and vascular patter of fingers [1], [7], [10], [16], [19], [22], [27], [44], [47], [48], [50], [54], [56], [57], [58], [63], [69], [70], [73], [76], [79], [98], [100], [101]. Also, some works provide combination of these biometrics, as well as different biometrics of the human body, e.g., eye iris and hand shape, palm print and face, voice and face [49], [58]. Systems applying combination of different biometrics are called multimodal biometrics systems.

Many hand geometric features can be measured and used to distinguish people, being these features unique to each person. Some of these measurements are: width, area, perimeter and thickness of the hand; length of the fingers; shape,

width and area of the region of the fingertips; and width at 1/3 and 2/3 of the fingers [69]. Also, combinations of these measurements such as the ratio of the length and the width can contribute to discriminate people.

On the feature matching or recognition phase, most of the systems compare the features of a test hand with the features of the hands which exist in a predefined database. This comparison is usually performed by some metric and a predefined threshold. By metrics, the distance between feature's vectors is calculated and if this distance is minor than the threshold, then the recognition or verification is accepted as true, otherwise it's rejected.

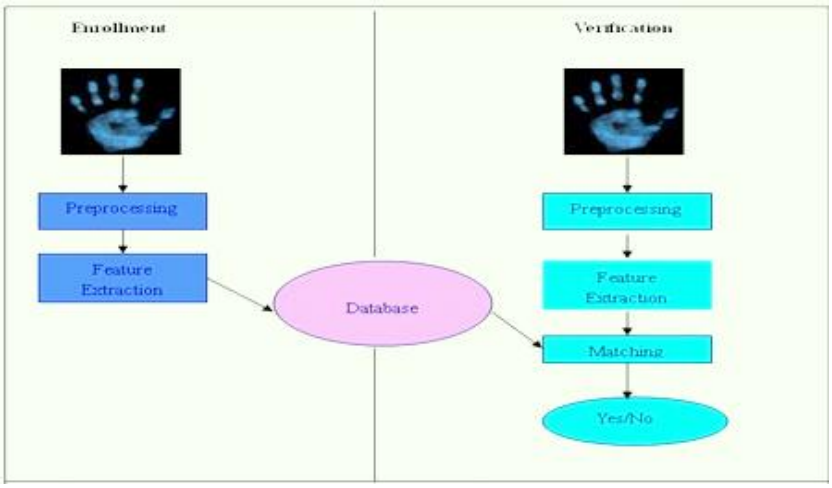

Figure 5.1- Representation of a hands biometrics recognition system [15].

The recognition systems commit two types of errors: one, in presence of "genuine hands", and the other in presence of "impostor hands". In the case of impostor hands, the distance between feature's vectors is minor than the threshold and so, the recognition is accepted as true although being false, being then a false

acceptance. In the presence of genuine hands, the distance between feature's vectors is greater than the threshold and so, the recognition is rejected although being true, it's a false rejection. There is a trade-off between the false rejection rate (FRR) and the false acceptance rate (FAR) in biometric systems. As both FRR and FAR are functions of the threshold, if the threshold decreases to make the system more secure, then the FRR increases. On the other hand, if the threshold rises to make the system more tolerant to input variations, then the FAR increases. The performance of the recognition systems can be analysed considering several thresholds and a graphic representation of the Receiver Operating Characteristic (ROC) curve. A ROC curve is a plot of FRR against FAR for various threshold values [49].

In the evaluation of the systems performance, the usual procedure consists on observing the ROC curve, and the intention is to choose a threshold value in which makes both FAR and FRR as smaller, as possible. Both these measures, as the correct recognition rate, can inform about the accuracy of a recognition system. Another measure used by some authors is the Equal Error Rate (EER) meaning the rate at which both FAR and FRR coincide. The lower EER corresponds to a system more accurate.

The recognition systems based on hand geometry are among the oldest methods used for automatic identification of people. One of the first known was used for security checks in Wall Street [30].

Next, we review the most referenced approaches in literature, addressing the hands biometrics for recognition.

Since 1971 several US patents devised mechanisms to validate or authenticate identification cards, such as credit cards. As in [24], it is presented a system providing cards with the hand geometry measurements. The measures, such as width and size of the hands are obtained manually by devices in the form of a box where the hands are inserted. These measures are encoded and recorded on the cards. Thus, at the time of presenting the card at a checkpoint, it is verified if the measures of the user's hand correspond to the measures encoded on his card allowing, this way, to validate or deny the authenticity of the card's user identity. Also, in [99] (1972), there is an automated system in which measures, by electromechanical means, predetermined

geometry measurements of the hands, including the distances between fingertips and finger lengths. A hand platform requires the palm of the hand and retains the hand and the fingers in a fixed position. Circuit means are provided automatically comparing the measures of the gauged hand with the correspondent measure of a selected hand that has been previously recorded. Depending on the presence or absence of the required correlation in the comparison within acceptable tolerance limits, the person's identity is confirmed or rejected, respectively. According to a certain tolerance limit, the best performance was about 99% of correct acceptance. Later another US patent work, at 1977 [23], uses the palm print biometrics which provide an apparatus that identifies a person based upon on the spacing of, at least, two preselected lines of the hand palm. For the recognition, it is compared the pattern of the palm lines with the pattern of the master palm line stored in computer. Another measures under consideration are the circular arc positioned between two fingers and the palm contour. Using bimodal biometrics, the work in [89] (at 1980), is based on hand shape and the hand geometry. This system consists of a palm pattern detector converting the hand palm and the palm contour into a bit pattern. Then, the palm pattern is corresponded to a number of binary bit data. Another involved feature is the palm convex part which is used as a parameter. Furthermore, the information concerning features of the five fingers, including the shape of the fingers tip, the joint region of the fingers and the length of the contour line of the finger tips is used. An individual is correctly identified in the identifying operation if all his features parameter coincide with the parameters of the correspondent features of an individual registered. An architecture using the three-dimensional hand information and a digital camera was created by another patent work at 1988 [80]. This apparatus consists of a digital camera and an optical measuring platen allowing a plan view and a side view of the hand. The operation is started by entering an identity code through a pushbutton keypad. The hand is then placed upon the measuring platen, and a three dimensional view of the hand is acquired by the digital camera. The geometry features extracted by the hands image are: lengths and widths of the finger; and thickness, area and perimeter of the hand. These features are then compared with the previously acquired and stored features allowing to confirm (or not) if the identity is the true identity. Three-dimensional apparatus like these also were used by the Recognition Systems, Inc. (RSI), Campbell, California, on the occasion of the Olympic Games of Atlanta at 1996 [36]. The prior characteristics of the athlete's hands or other enrolled Olympic personnel were registered in a database. Upon arrival

at the games, they placed their hands on a device for a three-dimensional geometry scan of the hand size and shape. The personnel entrance was permitted into the security area if there was a correspondence between the hand registered and the hand scanned.

Introducing digital cameras on the acquisition of hands images, several systems used pegs on the device board to guide the hand placement on the device. The works in [48] (at 1999) are the pioneers of recognition systems using pegs. These systems use five pegs on the hand acquisition image (see Figure 5.2) which apart from guiding the hand placement, are used to measure hand geometry features. They acquired 500 images from 50 people, 10 images per person. The features are 16, including the length and width of the fingers, a ratio of the palm (or palm and fingers) and the thickness of the hand. Some of these measurements are illustrated in Figure 5.3. The hand is represented as a vector of sixteen measures. The verification phase represents the process of comparing the currently acquired hand image with the one that is already in the database. The verification provides a positive result if the distance between both vectors is smaller than a defined threshold. The results obtained in terms of FAR were about 5%, and in terms of correct identification was 94.99%.

At the same year (1999), the work in [47], provides a system based on hands shape. The authors also use pegs to put the hand in a determined position but, unlike the previous work, the pegs are removed for the feature extraction. The hands images were taken from 53 people obtaining 353 hand images. For each person, the number of images taken is in the range [2, 15]. A contours algorithm is used to compute the hand shape. The five fingers of the hand are aligned according to a set of defined points. Each alignment produces a set of correspondent points. Given two hands images, the distances between the correspondent points are computed. The average of these distances considering all correspondence points defines the mean distance between two hands - Mean Alignment Error (MAE). At the verification stage, a pair of hands is identified as belonging to the same person if the MAE value is lower than the threshold. According to a determined threshold, the performance in terms of correct identification is 96.5% and FAR = 2%.

Figure 5.2- An example of a system with pegs [79].

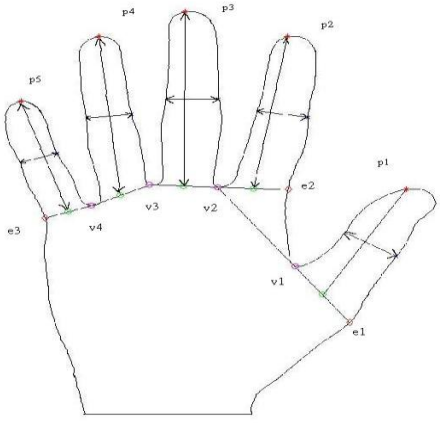

Figure 5.3- Some biometrics of the hand geometry [15].

The systems based on hand geometry usually assign features extracted from a hand to a vector. Some metric distance is applied to compute the similarity between features vectors which, one corresponds to the hand sample and the other corresponds to the hand of the database. Unlike these systems, some verification systems use probabilistic and machine learning techniques allowing classify hand sample. The most used techniques are, Support Vector Machines (SVM), k-Nearest Neighbour (KNN) and Gaussian Mixture Model (GMM) [7], [16]. Systems based on classifiers are trained for each of the enrolled people, classifying samples from other enrolled people, as the works in [79] (2000). These authors, like the ones in [48], use pegs to take the image and to extract the features. 31 hand geometry features are measured, such as widths and heights of the fingers and palm, angles between fingers and the horizontal. Some of the features were used in [48]. A statistical analysis is performed, determining the discriminatory features and allowing to reduce the features number to 25. The database has 200 images from 20 people. Regarding the verification phase, the recognition is based on similarities between feature's vectors using Euclidean distance, Hamming distance and GMM. A set of feature's vectors from the users enrolled in the system are trained for each output correspond to a class. Then, a new feature's vector is inputted, and if it's classified as one of the class in the database, means that there is a match between the hands. The best performance is considering GMM approach providing 97% of verification accuracy.

Hand geometry has been contact-based since the beginning of its use until now and can be classified as; 1) constrained or 2) unconstrained. While the first category requires a flat platform and pegs or pins to restrict the degree of hand freedom, in the second, the hands are free from pegs and pins, although still requires a platform to place the hand (e.g. scanner). There are more researches using the first category than the ones using the second, even though it gives the users more freedom in the process of image acquisition. This step is considered as the milestone from constrained contact-based systems [7].

A different approach, considering pegs free system or unconstrained contact-based, was proposed in 2001 [73] by a method to recognize hand shape through implicit polynomials. Fourth degree polynomials are used to model the fingers shape by a fitting curve. The features are the coefficients of the polynomial, combined with 16

geometry features. In this work were taken 30 hands images from 28 people by a digital camera without pegs, and a back lighted area to place the hands was imposed. From 30 images taken, 20 are for training and the remaining for testing. The recognition is based on the Mahalanobis distance. This procedure achieves 95% of performance and FAR = 1%.

The system developed in (2003) [57] presents great recognition rate, in which images were taken by a digital camera from 100 people, 10 per person. From these hands images, 160 combined characteristics are extracted from geometry and palm print: 16 features are from hand geometry and 144 features are from palm print. This is one of the first systems of person recognition using multimodal biometrics as palm print and hand geometry. The multimodal biometrics are combined by information fusion strategies, where the feature's vectors of the biometrics are concatenated to form a combined feature's vector. A similarity measure between the feature's vector from a user and the feature's vector from an identity is used as the matching score. This measure calculates the normalized correlation between two vectors, in which by a specified threshold, one can conclude about the recognition. By 10 hand's images collected from each user, 5 are used for training and the remaining for testing. The results for 472 test images, achieved 98.69% as recognition rate, FAR = 5.08% and FRR = 2.25%.

At the next work, unlike the previous ones, the hands images were captured by a scanner. In [10] (2004), the geometry features extracted are the same as in [79]. There are 714 images from right hand taken from 70 people by a scanner without pegs. The verification is based on the Chebyshev metric between feature's vectors. For each person, it is used a small number of hand's images as training set. Given a query feature's vector, the distances to feature's vectors derived of training hands images are measured. For verification, it is used a determined threshold to decide whether the feature's vectors are close enough to a given hand of training set. For a certain threshold, this system conducts to 98.5% of recognition performance and FAR less than 1%.

Applying the Independent Component Analysis (ICA) on the binary hand images to obtain the biometrics of hand shape, constitutes the researches in [54] (2006). Each hand's image is a combination of N sources of pixels or an N-dimensional

feature's vector. Due to the high dimensionality of the pixels from an image, there is a prior reduction stage by the statistical analysis, PCA (Principal Component Analysis). For the recognition stage, a hand of test is projected onto the set of predetermined sources of pixels and the result vector is compared with each N-dimensional feature's vector. The recognition occurs for the closest vectors according to a metric. The database consists of 1374 right hand's images from 458 people being 3 images of each right hand. The features extracted after the PCA application, are 271 features for each hand. The verification performance is about 98.21%.

Another research in [58] (2006), applies data reduction as also feature reduction. It analyses multimodal hands biometrics, using a single hand image taken from a digital camera with pegs-free to obtain features from the palm print and the hand shape. The palm print features are derived from discrete cosine coefficients by application of Discrete Cosine Transform (DCT). It allows transform hands images processing, and then data reduction. Regarding the hand shape, the features are seven. Also are included 16 hand geometry features. Considering the set of all the features, it is selected a feature subset (with a small number of features) intending to achieve similar or better performance than by using all the features. For that , it is applied the Correlation-based Feature Selection (CFS) classifier algorithm which uses an objective function based on correlation to evaluate the usefulness of all the features. Essentially, the idea is that the best feature subset must have high correlation with the class label but remain uncorrelated among them. During the application of the algorithm, the search is aborted if the addition of new features does not show any improvement in relation to the last best. The recognition phase is by some classifier algorithms, as SVM. The images of the database were collected from 100 people, consisting of 1000 images, ten images per person. The feature's vectors have 23 data from the hand shape and 144 data from the palm print image. Initially, these features are extracted for the feature evaluation and selection for the training data, which is constituted by 5 hand images from each person. One conclusion of this work is that, feature selection may reduce 52.08% the number of features, improving or maintaining the performance. The best performance achieves 98% of personal recognition.

One of the first works allowing to the user more freedom on the process of image acquisition is in [1] (2008). Also is proposed feature reduction, as the work

above. This system is based on using Natural Reference System (NRS) defined on the hands natural layout, in which neither hand pose nor prefixed position is required on the image acquisition process. The hands images were derived by scanner of the right and left hands, thus allowing to measure distances of the features for directly and crossed hands, as right/left, right/right, left/right, left/left. There are 5640 hand's images taken from 470 users. According to NRS, the contour of the hand is obtained and it is used to define the feature's vector. Initially, the feature's vector has the pixels belonging to the hand's contour. Then a polar representation of this vector provides some geometry features. Also, the correlations on a set of features are analysed, in which, features having high correlations are removed. The final feature's vector has 14 features. The performance evaluation is, 97.6% of correct identification, FAR = 1.3% and FRR = 1.3%.

Using graph representation for the feature extraction in [76] (2008), it is presented a biometric system based on new hand geometries. The image acquisition is made by scanner with fingers together and without pegs. During the image processing are detected 4 points at the top of the fingers (except thumb) and 2 points at the root of the 2 fingers, the forefinger and the little finger. These 6 points, define the vertex on the graph representation. The edges of the complete graph, which are 15, are the features extracted. In the verification process, the features of a test hand are compared with all stored patterns in the database. According to the distances from the test hand, are selected 3 hands from the database as the nearest neighbour of it. If one of them match with the test hand, then the person is verified, otherwise is rejected. The images were taken from 250 people and from each one 3 images. This system provides 99.11% of total success rate, FAR=2.97% and FRR=0%.

Analysing the system performance by taking into account differences on image resolution and also considering as the works in [1], more freedom in the process of image acquisition, are the researches in [27] (2009). The authors analyse the effect of changing the hand's image resolution over a hand geometry biometric system. They consider different image resolutions, from 120 dpi (the initial) to 24 dpi. The experiments are 4, whereas performed with 2 databases and 2 classifiers. The first database, acquires the hands images underneath and the second database, acquires the images over the hands. For the first database, the images are provided by a

scanner in which the users can place the hand freely. At the second database the images are taken by a webcam and the hands are placed on a white surface with several pegs. Both databases have hands images of 85 people and 10 different images of right hand of each one. The features are the width and height of each finger and the width of each finger at 70% of its height. Then, there are 3 measures for each finger and a total of 15 features extracted. The first classifier identifies by a multiclass SVM and the second classifier identifies by a neural network. For both classifiers, each database is divided into two databases, training and testing databases. The training database contains 4 images of each person and the testing database the remaining images. By these experiments, they concluded that an image resolution of 72 dpi offers the best performance. Also, they achieved an average recognition rate equals to 99.85% (with standard deviation 0.42), considering the first database, having more freedom on hand placement, as well the SVM classifier.

Another contribution also uses SVM classifier, but unlike the previous work the features are obtained of the palm print and hand geometry, by image segmentation. In this bimodal biometrics hand system, [98] (2009), the acquisition of the hands image is considering, fixed light source, black background and it is imposed that the fingers should be stretched and separated. At the pre-processed image phase, the images are converted to gray-level. The image of a palm is segmented and by the concept of Voronoi diagram it is cut into several blocks. In these blocks are fused the palm print and the hands geometry. The features extracted, are the statistic measurements as mean, of the gray level in each block. The hands images are, 1560 from 260 people, 6 images per person. Half of these 6 images are used to train a SVM allowing classify these images into 2 classes. In the recognition phase by the SVM classifier, the feature's vector is classified at one class, as the argued user, or it is classified at the other class, not alleged by the user. The performance of this system is described by, FAR = 0.0035% and FRR = 5.8%.

Addressing the problem of large pose hands variations, the method in [50] (2010) considers the biometrics of the palm print and the hand geometry. After being pre-processed the hands images in 3D space, the orientation and the normalization of the hands pose are estimated. The normalization of the hands images, allows extracting 3D points from a circular region around the centre of the palm by fitting a 3D

plan to these points. The normal vector of this plan is used to estimate the hands orientation. The plan and the normal vector, allows to estimate a set of 3D points which represents the corrected pose of the hand. These points, as well as the values corresponding to the intensity of the hand's image are further processed to locate regions of interest (ROI) for feature extraction. The recognition is by a distance between the feature's vectors computed by the Hamming distance. The hands images were taken by a commercially 3D digitizer from 114 people, 10 right hand's images per person, in which, 5 of them are for training. A hand sample is matched to all the remaining samples of the training data's user and the best match score is considered as the final score. This procedure is repeated for all 5 hand user's samples. The performance is provided by EER = 0.71%.

In a recent work (2011), it is presented a hand geometry biometric system for contact-less and platform-free scenarios in [16]. This system provides a template based on hand geometry distances, requiring information from only one individual, without considering the data from the rest of the individuals within the database. In the features extraction, it is considered measures of hand which are invariant to changes, like distance to camera, hand rotation and hand pose. The features are extracted by dividing the fingers from the basis to the tip into several parts. In each one of these parts, it is measured the width of fingers, based on the Euclidean distance between two pixels. And so, it is created a template collecting global information from samples of the same individual. This template proposes a matching method by minimizing the intra-class similarity and maximizing the inter-class likeliness. About the hands database, three different databases containing different acquisition procedures in respect to population size, distance to the camera, different illumination and hand rotation are enforced. The best results were considering the database which presents less variability in terms of; hand's rotation; distance to camera and environmental conditions, achieving EER = 1.4%.

The last papers don't clarify the percentage of correct identification.

Over the time, the researchers on hand's biometrics for recognition have been searching for a better performance of their templates in terms of recognition rate, FAR, FRR and EER. Also, increasing the population size, having more liberty on acquisition

of hands images, as well as evaluating the model by different classifiers, are factors which the researches look for take into consideration.

Time path of biometrics for recognition in literature

The subject of biometrics for recognition, as in particular the hand's biometrics for recognition, has got produced an increasing interest in the researchers. It is observed by the quantity of publications in the scientific journals, such as articles, conference paper, and book chapters, addressing this subject over the time. The graph representation in Figure 5.4 demonstrates it. The sources of this information are IEEE and Scopus libraries. The search was by title and abstract of the documents, considering the keywords "Biometric recognition" and after, refining this search by "hand". One can observe that, in the 5 years after the "2001, September 11", the hand's biometrics for recognition as the biometrics for recognition increased almost 10 fold the number of published papers (see, Figure 5.4).

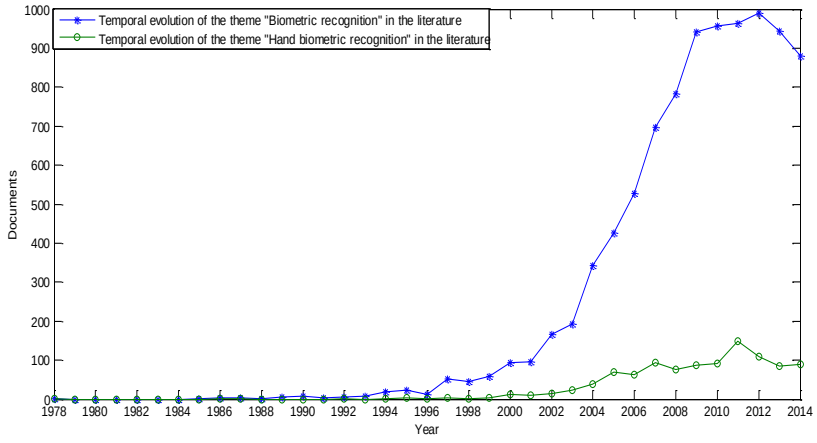

Figure 5.4- The evolution of documents published over the years covering the issues "biometric recognition" and "hand's biometric recognition".

5.4 Conclusions

The field of the biometrics technology addresses the automatic identification of people. Some personal attributes are used for biometric identification by recognition systems. These systems, as well as the systems based on the hand's biometrics have been developed in recent decades. In fact, an increasing interest by these issues is shown in Figure 5.4 based on the number of published papers. Noting that, in the period 2001-2008, the number of publications increased almost 10 fold. Furthermore, the identification by the hand has great potential because in literature it is pointed an average success rate of 97.6% (with 1.6% of standard deviation).

In this Chapter, we addressed the context of biometrics for recognition. We intend to use the biometrics of the hand's shape according to the works in [54]. These works, achieved 98.21% of success rate, which is more than the average success rate in literature. To the data sets constituted by these hand's biometrics, we propose us to apply the hierarchical clustering algorithms for the personal recognition (in Chapter 6) and to apply the consensus clustering techniques for the parental recognition (in Chapter 7).

Chapter 6

Comparative analysis of hierarchical clustering algorithms

6.1 Summary

The main goal of this Chapter is to compare the performances of the traditional hierarchical clustering algorithms and the approach SEP/COP. In these studies, we also include the application of both methodologies to a real world problem, particularly the hand's biometrics for recognition. The hierarchical clustering algorithms produce a nested set of partitions represented in a hierarchy. In the post-processing of the hierarchy, the partitions are defined by different levels of the hierarchy. A different post-processing of hierarchy is presented, called SEP/COP.

6.2 Introduction

In these studies we provide some comparative studies between the traditional hierarchical clustering algorithms addressed in Chapter 2, and the approach SEP/COP. For that, we include the use of the real data sets derived from the biometrics of hands for people's recognition. The results of the recognition rates of both approaches are obtained by the ARI values. These approaches are compared between them, as well as compared with the results of the literature.

A hierarchical clustering algorithm applied to a data set produces a series of nested partitions represented in a hierarchy. A hierarchy is a complex and difficult structure to interpret, so that, it is usual to post-process a hierarchy to find the best partition in it. In traditional approaches, each partition is defined by a horizontal line cutting the dendrogram at a determined level. The usual hierarchy's post-processing in some cases may not achieve the correct partition. So, the approach SEP/COP, to produce the correct partition has another procedure about the usual post-processing derived by the traditional hierarchical clustering algorithms.

6.3 The SEP/COP approach

In [37] is proposed a new method to obtain the best partition based on a wide set of partitions derived by a hierarchy. This method, called SEP (Search over Extended Partition set), looks for the best partition efficiently in a set of extended partitions. Finding the best partition on this set of partitions necessarily leads to results better or equal to that found in the set of partitions defined by the successive levels of the hierarchy, since all the extended partitions include the set of partitions provided by the hierarchy [5], [37].

The particularities of SEP algorithm restrict the use of validity indices, i.e., most of the available indices in literature cannot be used for extended partitions. So, the authors propose a new index of validity of clusters, called COP (whose acronym derives from the fact that check the properties of "Context-independent Optimality" and "Partiality").

The SEP/COP method is combined with the traditional hierarchical methods and deviates from those methods in which the partition is defined by a horizontal line cutting the dendrogram. The formal description of SEP/COP is as follows.

Let:

- X the data set to classify;
- P^Y a partial partition of X, as in Equation 6.1;
- $H = \{P_1, \dots, P_R\}$ a hierarchy of partitions of X, verifying the Equation 6.2;
- E_H, the set of extended partitions of the hierarchy, and T is the set of partitions built with combinations of clusters found in the hierarchy (see Equation 6.3).

$$P^Y = \left\{ C_1, \dots, C_k : \bigcup_{i=1}^{k} C_i = Y, C_i \cap C_j = \emptyset, \forall i \neq j, Y \subseteq X \right\} \tag{6.1}$$

$$\forall P_R, P_S \in H, R < S \Leftrightarrow \forall C_k \in P_R \; \exists C_l \in P_S : C_k \subseteq C_l \tag{6.2}$$

$$E_H = \left\{ P : P \subseteq T, \bigcup_{C \in P} C = X, \forall C_k, C_l \in P : C_k \cap C_l = \emptyset \right\}, \quad T = \bigcup_{C \in P, P \in H} C \tag{6.3}$$

Staring the dendrogram as a binary tree, the SEP method analyses each sub tree of the dendrogram independently and decides on each node, which one is the best partial partition to the data set.

The usual indices of validation of partitions cannot be applied to extended partitions, so, it is proposed a index of validation, called COP. This index is calculated by a weighted ratio of the intra-cluster variance and the inter-cluster variance of a partition, as in the Equation 6.4. The Equation 6.5 calculates the COP index of the union of two partial partitions. The lowest index value indicates the better partition, corresponding to the partition in which the clusters are more homogeneous and more separated between them.

$$COP(P^Y, X) = \frac{1}{|Y|} \sum_{C \in P^Y} |C| \frac{intra(C)}{inter(C)} \tag{6.4}$$

where,

$$intra(C) = \frac{1}{|C|} \sum_{X \in C} d(X, \bar{C}), \qquad inter(C) = \min_{x_i \notin C} \max_{x_j \in C} d(x_i, x_j)$$

$$COP(P^Y \cup P^Z, X) = \frac{1}{|Y| + |Z|} \left(\sum_{C \in P^Y} |C| \frac{intra(C)}{inter(C)} + \sum_{C \in P^Z} |C| \frac{intra(C)}{inter(C)} \right) =$$

$$= \frac{1}{|Y| + |Z|} (|Y| COP(P^Y) + |Z| COP(P^Z)) \tag{6.5}$$

$$0 \leq COP \leq 1$$

Description of the algorithm:

The idea of the algorithm is, first of all, to view the hierarchy as a tree with sub trees and inner nodes. Analysing each sub tree, at each node, it is decided which is the best partition between two partitions: one corresponds at the current node, and the other corresponds to each one of its child node. The comparison is made by the COP values and hence deciding for the best partition at each sub tree.

A demonstrative example of SEP/COP method procedure is represented on Figures 6.1 – 6.3 and Tables 6.1, 6.2. In Figures 6.1a) and 6.1b) the dark lines define the local partitions P^{Y_1} and P^{Y_2}, respectively, and the red line the partition P^{Y_3}. Comparing the COP values of these partitions and the unions of partitions, we have four hypotheses for the resulting local best partition. They are represented in Figures 6.1c), 6.1d), 6.1e) and 6.1f). The Table 6.1 reports these possible relations of COP values between the partitions and the consequent local best partitions.

Assuming that the best local partition is depicted on Figure 6.1d) and considering now the Figures 6.2 a) and 6.2 b) where the dark lines define the partitions P^{Y_4} and P^{Y_5} respectively and the red line the partition P^{Y_6}. If we compare COP values of these partitions and the unions, we have again four hypotheses for the resulting best partition represented in Figures 6.2c), 6.2d), 6.2e) and 6.2f). The Table 6.2 reports the possible relations of COP values and the consequent local best partitions.

Finally, Figure 6.3 illustrates some of the possible final partitions resultants of SEP/COP method. One can observe that it can be quite different of the partitions obtained by the traditional hierarchical clustering algorithms.

Table 6.1: The relations of COP values at the local partitions and the correspondent representative Figure of the best local partition.

Comparison the COP values of the partitions	Figure
$COP(P^{Y_1}, X) < COP(P^{Y_1} \cup P^{Y_3}, X) \wedge COP(P^{Y_2}, X) > COP(P^{Y_2} \cup P^{Y_3}, X)$	6.1c)
$COP(P^{Y_1}, X) > COP(P^{Y_1} \cup P^{Y_3}, X) \wedge COP(P^{Y_2}, X) < COP(P^{Y_2} \cup P^{Y_3}, X)$	6.1d)
$COP(P^{Y_1}, X) > COP(P^{Y_1} \cup P^{Y_3}, X) \wedge COP(P^{Y_2}, X) > COP(P^{Y_2} \cup P^{Y_3}, X)$	6.1e)
$COP(P^{Y_1}, X) < COP(P^{Y_1} \cup P^{Y_3}, X) \wedge COP(P^{Y_2}, X) < COP(P^{Y_2} \cup P^{Y_3}, X)$	6.1f)

Table 6.2: The relations of COP values at the local partitions and the correspondent representative Figure of the best local partition (continuation).

Comparison the COP values of the partitions	Figure
$COP(P^{Y_4}, X) < COP(P^{Y_4} \cup P^{Y_6}, X) \wedge COP(P^{Y_5}, X) < COP(P^{Y_5} \cup P^{Y_6}, X)$	6.2c)
$COP(P^{Y_4}, X) < COP(P^{Y_4} \cup P^{Y_6}, X) \wedge COP(P^{Y_5}, X) > COP(P^{Y_5} \cup P^{Y_6}, X)$	6.2d)
$COP(P^{Y_4}, X) > COP(P^{Y_4} \cup P^{Y_6}, X) \wedge COP(P^{Y_5}, X) < COP(P^{Y_5} \cup P^{Y_6}, X)$	6.2e)
$COP(P^{Y_4}, X) > COP(P^{Y_4} \cup P^{Y_6}, X) \wedge COP(P^{Y_5}, X) > COP(P^{Y_5} \cup P^{Y_6}, X)$	6.2f)

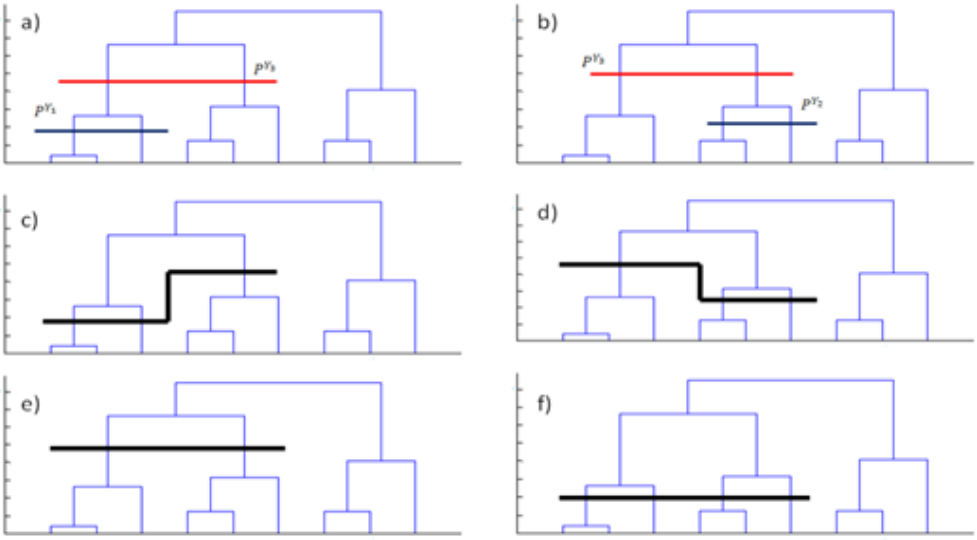

Figure 6.1- A demonstrative example on application of SEP/COP method in a hierarchy.

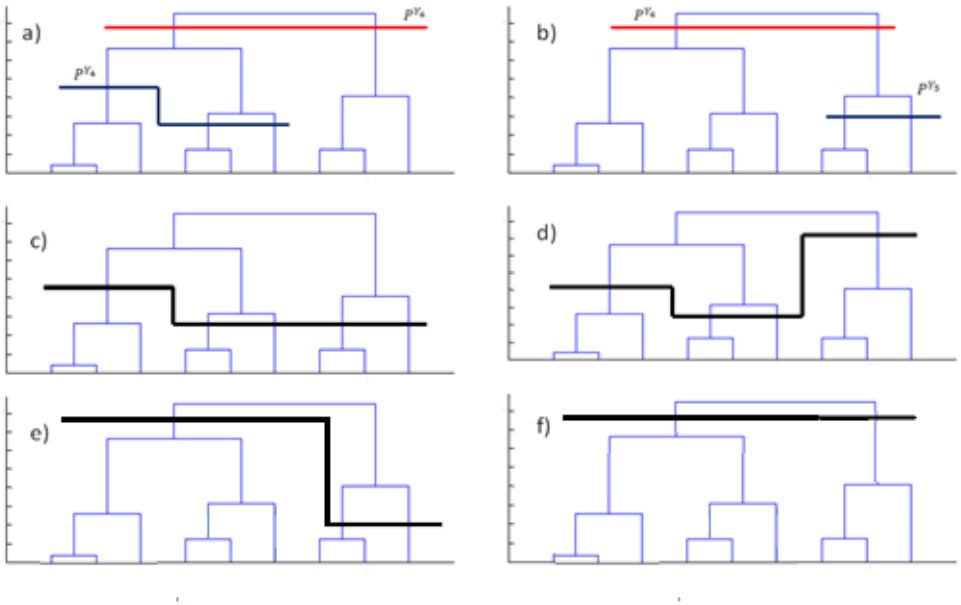

Figure 6.2- The sequel of the demonstrative example on application of SEP/COP method in a hierarchy.

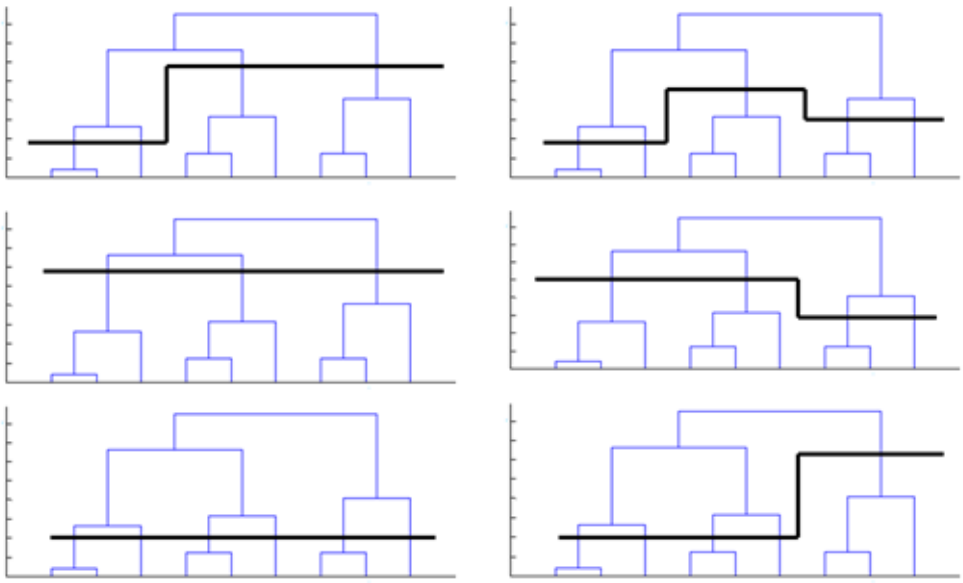

Figure 6.3- Some of the possible final partitions by the demonstrative example.

6.4 Experimental design

In these experiments, we apply the traditional hierarchical clustering algorithms, SL, CL and AL, where it is considered as measure of proximity the Euclidean distance. These algorithms showed in our experiments in Chapter 3 that they may have very different performances and variabilities between them when applied to a given data set. Also, each one of these algorithms may have very different performances and variabilities when applied to different data sets. These are the reasons why we consider the referred algorithms in the current studies.

In the nested partitions provided by the hierarchical algorithms, the clusterings are obtained by the partitions according to the number of clusters of the known data structure. Also, clusterings obtained by using the SEP/COP approach are considered, the combined method of finding the best extended partition.

In the evaluation of the resulting clusterings, it is applied the Adjusted Rand index (ARI) between the clusterings obtained and the true known clustering, by the external validation criterion.

Thus, we compare the traditional hierarchical clustering algorithms and the SEP/COP method by empirical studies using synthetic and real-world data sets. For the simulated data sets, different structures into clusters are considered. Also, it is analysed the stability of the solutions by disturbance, including noise in data. Concerning the real-world data set, it is a multidimensional data set with the hand's features for personal recognition.

According to the real-world data set, the experiments are performed over the features extracted from the hand's images of people. We perform our experiences on six sizes of selected population, namely, population subsets consisting of 20, 35, 50, 70, 100 and 458 people, so that we can compare our results with the results in the literature [54]. For each population subset, we apply the traditional hierarchical clustering and the SEP/COP algorithms to the data set with the features of the hands, obtaining then the clusterings. These clusterings are compared by the ARI, with the true clusterings. The true clusterings have in each cluster the features of the hand of each person.

For the simulated data sets, for each established structure into clusters it is generated 1000 data sets. To each one of these data set is applied the traditional hierarchical clustering and the SEP/COP algorithms, obtaining 1000 clusterings for each clustering algorithm. Each resultant clustering is compared with the known structure into clusters by the ARI. The average and the standard deviation of the ARI values are computed. Also, the number of times that the true clustering is achieved is counted, i.e., the number of times that the ARI is equals to 1.

It follows the description of the data sets.

6.4.1 Data sets

Simulated data sets

In order to reach the variety of situations regarding the data sets, we consider different data sets with respect to their clusters as the number of clusters and the respective cardinality, shape and homogeneity, clusters well separated and quite close.

The 2-dimensional simulated data sets used in our experiments are represented in Figures 6.4 - 6.8 and the details of these data sets are shown in Table 6.3. These data sets are with random data and Normal distribution (according to their structure into clusters). Some of them are data sets used in our experiments in Chapter 3. Also, in some data sets we introduce data noise, randomly, uniformly distributed, near to a cluster.

The data sets are with 3 and 10 clusters, with the general nomenclatures, d1c3 and d2c10, respectively. Regarding the data sets d1c3, they have two clusters equally homogeneous and near to each other, while the remaining cluster is less homogeneous and apart from the others. We consider varying the cardinality of these clusters, considering three situations which are, clusters with different cardinalities, 10×50×50, clusters with the same cardinality, 20×20×20 and clusters with the same cardinality but with greater size, 50×50×50. Furthermore, for each one of the three situations, relatively to the two closest clusters, we consider two scenarios: making these clusters the closest or make them a bit apart. Lastly, for some data sets, different levels of data noise are introduced as 4% and 10% of new elements to be clustered. Regarding the data sets, d2c10, having ten clusters, we also consider varying the homogeneity, separability and the cardinality of the clusters. Each cluster has the mean value randomly in range (0,50), the variances in range (0.1,3) and the number of elements of each cluster in range (24,51). Each cluster is constructed by imposing conditions avoiding overlapping clusters and ensuring that no cluster is too close to another cluster. Also, it is introduced data noise, namely at, 5%, 10% and 20%.

For each situation described above, 1000 data sets are constructed.

Real-world data set

The real-world data set is derived from the hand's images database and the data are the features of these images, available at the Bosphorus Hand Database [9]. This database consists of the right hands images from 642 people, 3 hands images per person. From each image, 271 features are extracted. Those features are based on the shape of the hand researched in [54] where, in the recognition stage, the feature's hand vector of test is compared to a set of feature's vectors existing in a data set. The recognition occurs for the closest vectors according to a metric. In our experiments through the clusterings, the recognition occurs if the feature's vectors of the 3 hand's images of a person are all together in the same cluster, moreover, if each cluster only contains 3 feature's vectors of a person.

Table 6.3: Details of the simulated data sets. Data generated by Normal distribution, $N(\mu\,, \sigma^2)$ where μ is the mean and σ^2 is the variance. C is the number of clusters, Ni is the number of data of the cluster i and AN is the noise added. The data noise are generated by Uniform distribution U(a,b) where (a,b) is the support interval.

Data set	C	Ni	Source	AN	
d1c3v1_1	3	50×50×50	C1: $\mu_x = -1, \mu_y = 0, {\sigma^2}_x = {\sigma^2}_y = 0.3$ C2: $\boldsymbol{\mu_x = 1.5, \mu_y = 2.5}, {\sigma^2}_x = {\sigma^2}_y = 0.3$ C3: $\mu_x = 8.5, \mu_y = 10, {\sigma^2}_x = 1.5, {\sigma^2}_y = 2.25$	No	
d1c3v1_2		20×20×20			
d1c3v1_3		10×50×50			
d1c3v1_1n4		50×56×50		4% : U(3,4)	
d1c3v1_1n10		50×56×59		10% : U(3,4) ×U(6,7)	
d1c3v2_1		50×50×50	C1 : $\mu_x = -1, \mu_y = 0, {\sigma^2}_x = {\sigma^2}_y = 0.3$ C2: $\boldsymbol{\mu_x = 2.5, \mu_y = 2.5}, {\sigma^2}_x = {\sigma^2}_y = 0.3$ C3: $\mu_x = 8.5, \mu_y = 10, {\sigma^2}_x = 1.5, {\sigma^2}_y = 2.25$	No	
d1c3v2_2		20×20×20			
d1c3v2_3		10×50×50			
d2c10	10	Random in [25,50]	Ci: $\mu_x, \mu_y \in [0, 50]$; ${\sigma^2}_x = {\sigma^2}_y \in [0.1, 0.3]$, i=1,..10. For each 2 clusters, $d(\mathrm{C_k}, \mathrm{C_l}) > 3(\sigma_k + \sigma_l)$ where $\mathrm{C_k}$ and $\mathrm{C_l}$ are the centre points and σ_k and σ_l are the standard deviations, respectively	No	
d2c10n5				$X: d(\mathrm{C_1}, X) = \sigma_1 + \mathrm{U}(0,1)$, where $\mathrm{C_1}$ is the centre point and σ_1 is the standard deviation of the cluster C1	5%
d3c10n10					10%
d3c10n20					20%

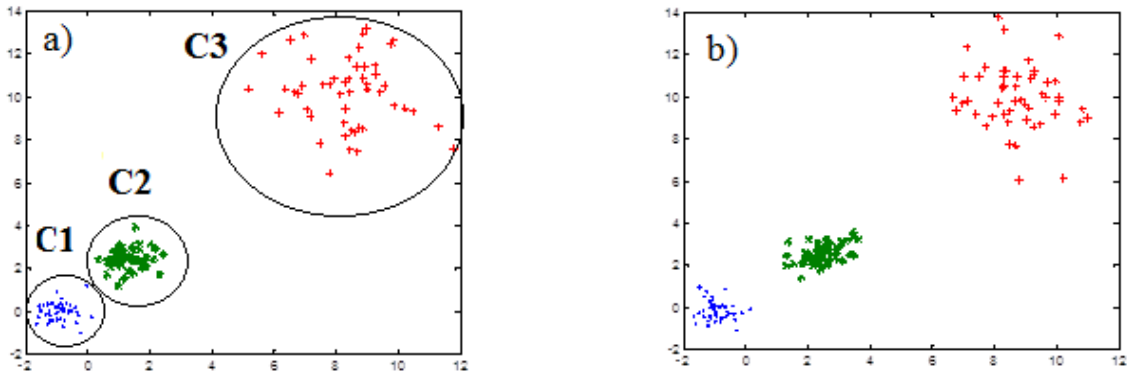

Figure 6.4- Representation of the data sets **a)** d1c3v1_1, **b)** d1c3v2_1 and clusters C1, C2, and C3.

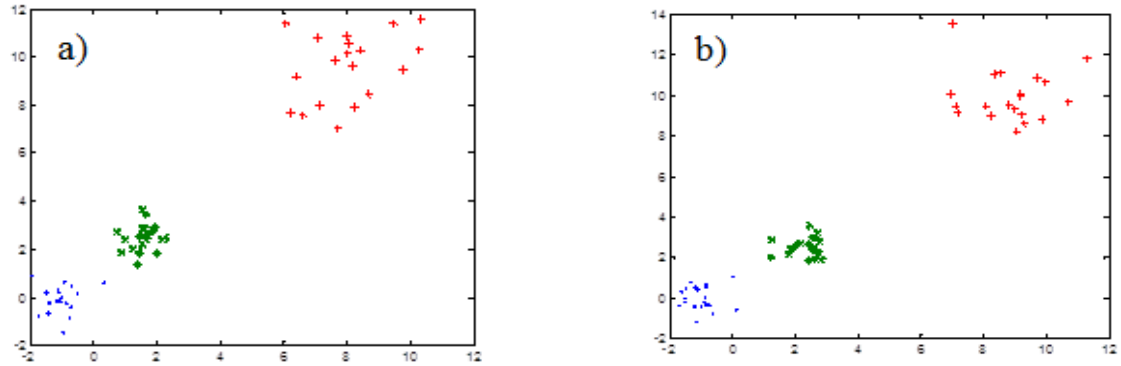

Figure 6.5- Representation of the data sets, **a)** d1c3v1_2, **b)** d1c3v2_2.

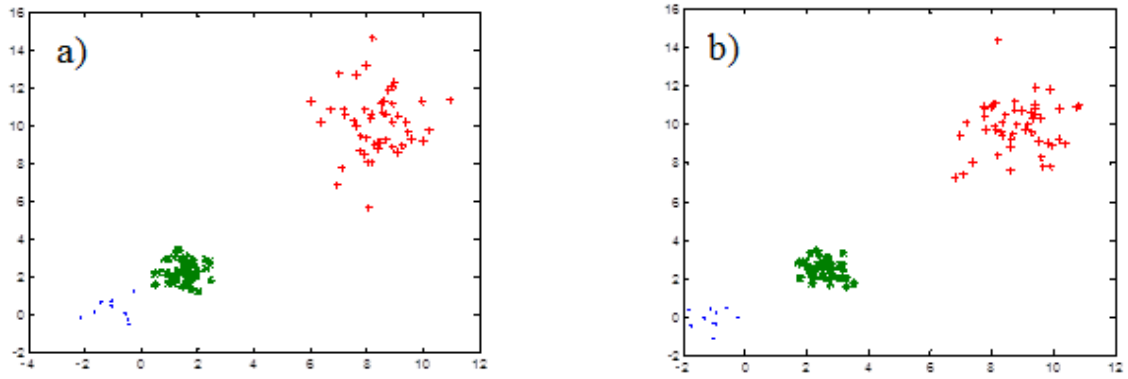

Figure 6.6- Representation of the data sets, **a)** d1c3v1_3, **b)** d1c3v2_3.

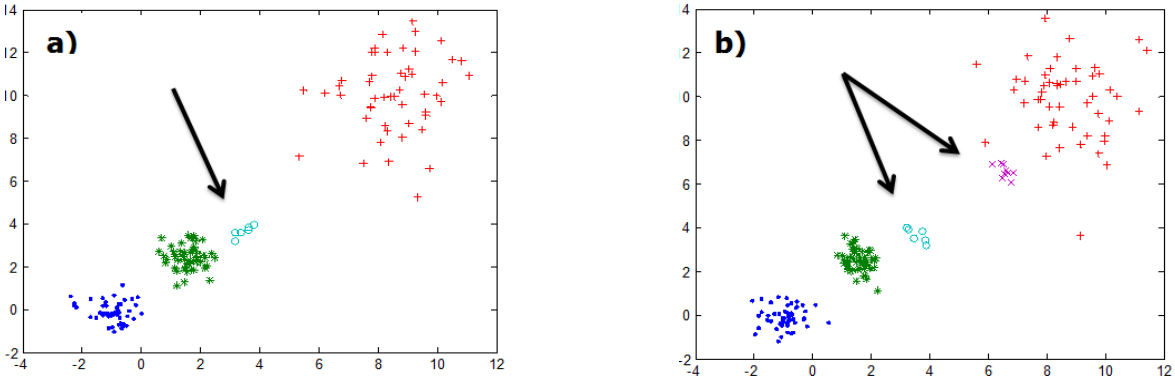

Figure 6.7- Representation of the data sets, **a)** d1c3v1_1n4, **b)** d1c3v1_1n10, with noise data marked by arrows.

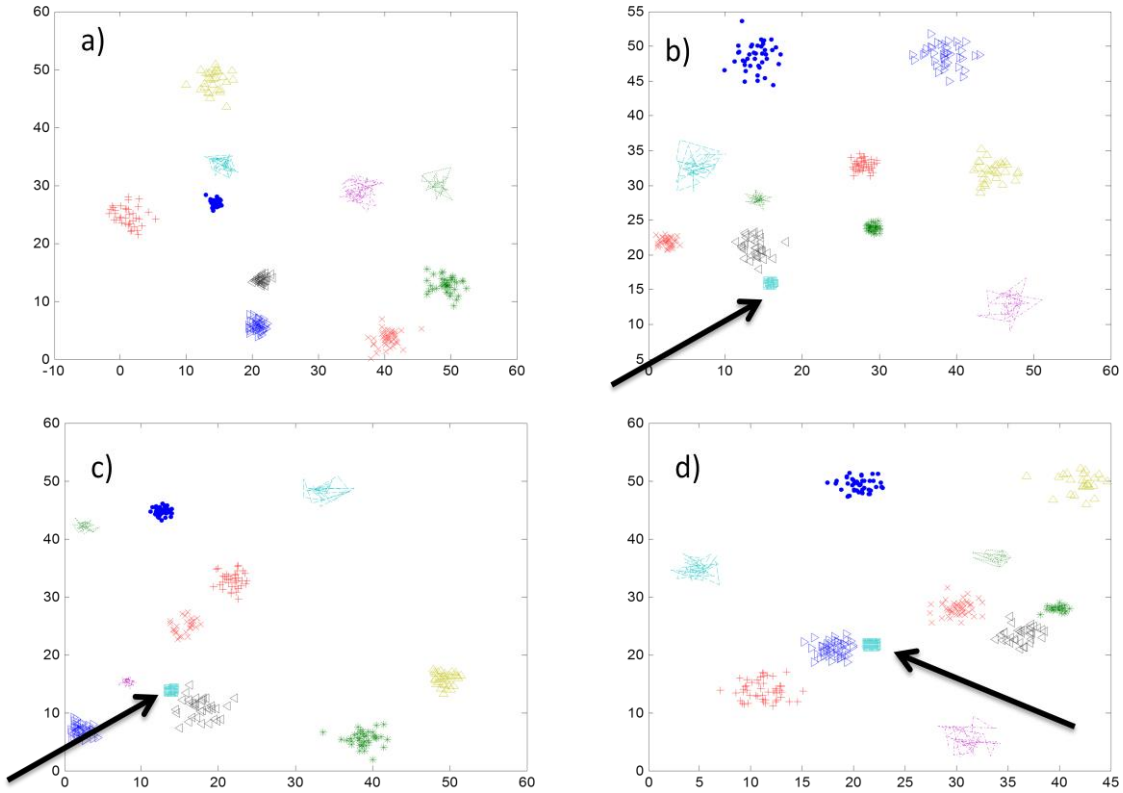

Figure 6.8- Representation of the data sets d2c10 with different noise levels, marked by the arrows, **a)** without noise, **b)**,**c)**, **d)** with 5%, 10% and 20% of data noise, respectively.

6.4.2 Results and discussion

Starting by the analysis of the results about the simulated data sets, namely those showed in Table 6.4, we conclude that:

- The data sets in accordance with their clusterings having clusters with the same cardinality even different homogeneities, as d1c3v1_1 and d1c3v1_2, the traditional algorithms, SL and AL, outperform the SEP/COP (higher, ARI average and recovery rate). But when the separability of the nearest and more compacts clusters increases, as in d1c3v2_1 and d1c3v2_2 respectively, all the algorithms improve but the SEP/COP outperforms the traditional algorithms (higher, ARI average and recovery rate and smaller standard deviation). Moreover, this approach presents good performances (ARI averages over 0.99 and the most of recovery rates over 98%).
- Regarding the clusterings having clusters with different cardinalities, homogeneity and separability, as d1c3v1_3 and d1c3v2_3, SL and AL algorithms outperform the SEP/COP (higher, ARI average and recovery rate). Regarding the data set, d1c3v1_3, some clusterings obtained by SEP/COP method, have a certain agreement with the true clustering, by the considerable ARI average obtained (over 0.83). But, none of these clustering is the true clustering, by the null recovery rate observed.
- Unlike the traditional algorithms, the results obtained by the SEP/COP approach, considering the different algorithms SL, CL or AL, are very close.
- The presence of data noise affects more the traditional than the SEP/COP algorithm. In fact, the performance of SEP/COP is even more apparent with the noise increasing (higher, ARI average and recovery rate).

Regarding the results in Table 6.5, for the simulated data sets in which the clusterings have clusters with random cardinality, homogeneity and separability. Some of these clusters can be close from another, but the majority of the clusters are well apart. Both the approaches have similar performance. Again, one can observe that the SEP/COP is less influenced by data noise than the traditional algorithms.

In summary, regarding the simulated data sets, with a natural data clustering, in which the clusters have the same cardinality, different homogeneity and having compact clusters close to each other, it is observed that the traditional algorithms have similar performance to the SEP/COP algorithm, in some cases even better. But, being these compact clusters a bit more separated, the SEP/COP produced better results than the traditional algorithms, achieving ARI average greater than 0.99. Furthermore, the SEP/COP presents a good performance at all cases with presence of noise and has similar performances using different aggregation methods, as SL, CL or AL.

Now, analysing the results about the real-world data set, shown in Table 6.6, according to the ARI values, we note that the SEP/COP approach almost always achieves the higher values in relation to the traditional algorithms. This approach achieves the best ARI value (equals to 1), by applying AL algorithm to the data sets according to the sizes, 20, 35, and 50 people. Also, achieves ARI value, 0.99, for database of 100 people.

As ARI is a measure based on agreements between two clusterings at the context of hand's biometrics for recognition, we consider that this measure provides the correct percentage of identification of people. So, in Table 6.7 are shown the best percentages of recognition achieved by the traditional hierarchical and SEP/COP algorithms, for different sizes of data set. Also, for comparison, is presented the results obtained in the literature [54]. The SEP/COP method, achieves 100% of correct identification for some data sets. This means that it is able to identify correctly all the people, namely for the data sets consisting of 20, 35 or 50 people, outperforming the works in the literature. Even the traditional hierarchical clusterings present 100% of recognition, for the data set of smaller size. When the data sets is scaled up to greater sizes the results show that the SEP/COP algorithm can handle with even larger data sets, with little bit degradation of performance (approximately greater or equal to 95% of identification) and still outperforming the works in the literature, according to the data set having 100 people.

Table 6.4: For each simulated data set, comparison between the traditional hierarchical clusterings and the SEP/COP algorithms in terms of: **A**- the average and the standard deviation of ARI and **B**- the percentage (in 1000) of recovery of the true clustering.

A

Data set		Traditional	SEP/COP
d1c3v1_1	SL	0.6660 (0.1915)	0.6307 (0.1521)
	CL	0.4959 (0.1205)	0.6307 (0.1521)
	AL	0.6982 (0.2148)	0.6299 (0.1513)
d1c3v2_1	SL	0.8898 (0.1914)	**0.9981 (0.0273)**
	CL	0.6116 (0.2361)	**0.9976 (0.0305)**
	AL	0.8843 (0.1952)	**0.9981 (0.0273)**
d1c3v1_2	SL	0.7266 (0.2306)	0.6578 (0.1802)
	CL	0.6114 (0.2391)	0.6569 (0.1796)
	AL	0.7737 (0.2399)	0.6578 (0.1802)
d1c3v2_2	SL	0.9141 (0.1804)	**0.9929 (0.0549)**
	CL	0.7655 (0.2645)	**0.9924 (0.0566)**
	AL	0.9268 (0.1701)	**0.9925 (0.0565)**
d1c3v1_3	SL	0.9070 (0.0932)	0.8332 (0)
	CL	0.6688 (0.0717)	0.8331 (0.0011)
	AL	0.8656 (0.0987)	0.8331 (0.0011)
d1c3v2_3	SL	**0.9755 (0.0626)**	0.8543 (0.0556)
	CL	0.7225 (0.1357)	0.8544 (0.0553)
	AL	**0.9544 (0.0815)**	0.8544 (0.0558)
d1c3v1_1n4	SL	0.6601 (0.1978)	0.7337 (0.2176)
	CL	0.7554 (0.2638)	0.7353 (0.2182)
	AL	0.7536 (0.2297)	0.7362 (0.2183)
d1c3v1_1n10	SL	0.6804 (0.1870)	**0.9458 (0.1360)**
	CL	0.5613 (0.1966)	**0.9567 (0.1242)**
	AL	0.5534 (0.1272)	**0.9551 (0.1262)**

B

Data set		Traditional	SEP/COP
d1c3v1_1	SL	24.7	14.5
	CL	4.6	14.5
	AL	31.2	14.3
d1c3v2_1	SL	75.1	**98.8**
	CL	26.4	**98.4**
	AL	73.7	**98.8**
d1c3v12	SL	41.4	21.7
	CL	26.5	21.5
	AL	51.7	21.7
d1c3v2_2	SL	81.5	**98.3**
	CL	55.2	**97.9**
	AL	84.1	**98.2**
d1c3v1_3	SL	49.9	0
	CL	1.8	0
	AL	33.4	0
d1c3v2_3	SL	**86.6**	12.7
	CL	16.7	12.2
	AL	**75.8**	12.8
d1c3v1_1n4	SL	25.0	38.9
	CL	49.5	39.6
	AL	44.1	39.9
d1c3v1_1n10	SL	25.1	**83.3**
	CL	15.5	**86.3**
	AL	6.4	**86.4**

Table 6.5: For each simulated data set, comparison between the traditional hierarchical clusterings and the SEP/COP algorithms in terms of the average and the standard deviation of ARI.

Data set		Traditional	SEP/COP
d2c10	SL	0.9825 (0.0390)	0.9826 (0.0368)
	CL	0.9873 (0.0401)	0.9896 (0.0279)
	AL	0.9886 (0.0361)	0.9885 (0.0275)
d2c10n5	SL	0.8530 (0.0828)	0.9306 (0.0467)
	CL	0.9102 (0.0549)	0.9024 (0.0719)
	AL	0.9066 (0.0357)	0.9024 (0.0719)
d2c10n10	SL	0.8628 (0.0748)	**0.8916 (0.0579)**
	CL	0.8616 (0.0746)	**0.8914 (0.0522)**
	AL	0.8608 (0.0750)	**0.8987 (0.0472)**
d2c10n20	SL	0.7362 (0.0517)	**0.8560 (0.0650)**
	CL	0.7490 (0.0427)	**0.8504 (0.0693)**
	AL	0.7468 (0.0460)	**0.8560 (0.0650)**

Table 6.6: For the real data sets, comparison between the traditional hierarchical and the SEP/COP algorithms in terms of ARI value for a given size of data set.

Size of data set		Traditional	SEP/COP
20	SL	0.9102	**1**
	CL	0.9102	**1**
	AL	**1**	**1**
35	SL	0.8656	0.9902
	CL	0.8997	**1**
	AL	0.9483	**1**
50	SL	0.8720	0.9932
	CL	0.8720	0.9796
	AL	0.8639	**1**
70	SL	0.8391	0.9424
	CL	0.9488	**0.9495**
	AL	0.9244	**0.9495**
100	SL	0.8286	0.9898
	CL	0.8729	0.9833
	AL	0.8565	**0.9916**
458	SL	0.3659	0.9493
	CL	0.7885	0.9457
	AL	0.7265	**0.9518**

Table 6.7: Comparison of the correct recognition percentage, by the best result of the traditional hierarchical and SEP/COP algorithms with the results in [54] for a given size of data set.

Size of data set	[54]	Traditional	SEP/COP
20	99.48	**100**	**100**
35	99.40	94.83	**100**
50	99.27	87.20	**100**
70	99.03	94.88	94.95
100	98.81	87.29	**99.16**
458	97.31	78.85	95.18

6.5 Conclusions

In this Chapter we focused on the problem of searching the best clustering in hierarchical algorithms. The procedure was made in the nested set of partitions, defined by the hierarchy. In the traditional approaches each partition is defined by a horizontal line cutting the hierarchy or dendrogram at a determined level. In [37] it is proposed an improved method SEP/COP, to obtain the best partition based on a wide set of partitions. This approach includes a proposed index of validity of partition adapted to this new situation. Being that, the best partition is achieved by this index instead of defined by cutting the dendrogram as the traditional algorithms.

Studies of traditional hierarchical clustering algorithms (addressed in Chapter 2) and the approach SEP/COP for choosing the best partition when interpreting a hierarchy, were performed in this Chapter.

For that, we did a comparative study of these two types of approaches through a set of experiments using two-dimensional synthetic and the real-world data sets. Regarding the simulated data, these experiences didn't allow to choose an approach since any approach has proved to be, at all situations, consistently better. The SEP/COP algorithms proved to be good solutions towards situations having in data clusterings, clusters well apart even homogeny and clusters with the same cardinality. Also, these algorithms are a bit dependent on the algorithm applied and more robust to the presence of data noise.

About the real world data set, related to the person's recognition systems, by the features extracted from the hands, the SEP/COP algorithms usually prove to have a better performance than the traditional ones. Furthermore, they attain a performance of 100% of correct identification for data sets with 20, 35 and 50 people. Also, they present 99% of correct identification for the data set with 100 people. These results

outperform the results in literature. So, the results of our experiments demonstrated that the SEP/COP algorithms can contribute to the identification systems based on the hand's biometrics.

Chapter 7

Comparative analysis of consensus clustering

7.1 Summary

In this Chapter we intend to compare the performance of some contributions to the consensus clustering. These contributions are some traditional approaches and a multi-objective consensus clustering techniques. The multi-objective technique allows to find more than one relevant structure that may exist in a data set. Regarding the real data set, a database of hand's images of parents and children is constructed to investigate if the consensus techniques are able to recognize the parents and their children.

7.2 Introduction

The goal of this Chapter is to perform some comparative studies on some consensus clustering techniques, namely the traditional addressed in Chapter 2 and a multi-objective MOCLE.

The traditional consensus clustering has some inherent problems. One difficulty to the traditional consensus clustering techniques is that they conduct to a single solution, wherein it's possible that a data set can have more than one relevant data structure. Moreover, the existence of individual clusterings with poor quality can influence negatively the quality of the consensus. Also, it is often necessary to pre establish the number of clusters of the consensus, which it is difficult by the data structure usually being unknown.

The multi-objective approach overcomes these difficulties providing, instead of a single structure, a set of alternative structures leading to different interpretations of the data which can be very helpful to the expert in the field [40].

7.3 Multi-objective consensus clustering

The most common clustering algorithms use only an objective function which allows obtaining a single structure, limiting the other knowledge that can be extracted from the data. The algorithms in which the clustering are obtained by multi-objective optimization have the intention to overcome this limitation since that deal simultaneously with more than one objective function, called multi-objective clustering algorithms.

One of the main multi-objective clustering algorithms is MOCK – Multi-Objective Clustering with automatic K-determination [40], which is able to find structures in

clustering with multi criteria and also determine the number of clusters. MOCK uses an evolutionary multi-objective algorithm, PESA II – Pareto Envelope based Selection Algorithm [13] and has two objective functions, which are compactness and connectivity. The evolutionary algorithms have been used in many works for being easily applied to the optimization problems, since they are based on Pareto optimization. The evolutionary algorithms simulate the natural evolution in a population, where there are individuals and genetic information. The idea is to keep a set of candidates' solutions which are manipulated by the genetic operators going by a selection process along the iterations. A Pareto optimal, in general means that it can't be improved for any objective without getting worse than another objective. The objective functions in the optimization criteria reflect the interest of have clusterings with quality. The compactness of the clusters is expressed by the intra-clusters variances. The connectivity reflects the degree in which neighbouring objects are placed in the same cluster in a clustering. Both objective functions are for to minimize in the optimization process.

The consensus clustering multi-objective is an approach resulting from the combination of multi-objective clustering algorithms and the traditional consensus clustering techniques.

By the fact that the multi-objective clustering algorithms can find many solutions, usually lead to the more difficult analysis by domain expert. A multi-objective consensus clustering is concerned to, multi-objective algorithms and consensus clustering techniques in the optimization process [25], [66]. The Multi- Objective Clustering Ensemble, MOCLE [26], applies an evolution process to the individual clusterings and pairs of the resulting clusterings are combined iteratively, by a consensus clustering technique, optimizing some established criteria. MOCLE starts with the generation of individual clusterings by the application of various clustering algorithms and different parameters to a data set. These individual clusterings are the initial population in the evolutionary algorithm based on Pareto, the genetic algorithm NSGA-II - Non-dominated Sorting Genetic Algorithm [17]. MOCLE uses this genetic algorithm only considering the crossover operator of the individuals. In the combination of pairs of clusterings, MOCLE uses a graph representation and MCLA algorithm. The

graph is partitioned into k parts, by METIS partitioning algorithm, being k the number of clusters of the resulting clustering of this combination and it is randomly chosen within the range of the number of clusters of the two combined clusterings. The resulting clustering of the combination is an individual of the population in the genetic algorithm. In MOCLE the two optimization criteria, compactness and connectivity, are defined by the objective functions in the genetic algorithm and are the validation indices. These functions represent the quality measures of a clustering.

Next, we proceed to some experiments applying the consensus clustering technique, MOCLE, and the traditional techniques discussed in Chapter 2.

7.4 Experimental design and results

At these experiences, we compare the performances of some consensus clustering techniques. For that, we apply the traditional consensus clustering techniques, as in our experiments in Chapter 4, namely, TEC.1 [31], TEC.2 [33], TEC.3 [87], [88] and also the multi-objective technique, TEC.4 [26]. Regarding the multi-objective technique we use the version of MOCLE available at the server laboratory of Intelligent and Distributed System, of the Federal University of São Carlos, Brazil. This technique, unlike the traditional ones, can provide more than one consensus solutions. Despite this, in the results we refer to the consensus clustering which presents the best performance (greater ARI value) in relation to the remaining ones.

We proceed to a series of experiments for the performance analysis and comparison of these different approaches. The evaluation of the consensus clustering obtained is performed by using the ARI (by the agreement between each consensus clustering and the known clustering).

The consensus clustering techniques are applied to sets of individual clusterings or base clusterings. For obtaining the individual clusterings, we consider the traditional hierarchical clustering algorithms, SL, CL, AL and W, because they can present very different performances between them unlike the SEP/COP approach.

As the hierarchical clustering algorithms provide a set of nested partitions, we consider the partition obtained by cutting the hierarchy on a determined level according to the number of clusters of the known clustering.

To each data set is applied these algorithms then obtaining 4 clusterings which are the base clusterings for the consensus techniques.

In the set of experiments are considered simulated data sets, namely D1-4g, D2-3g, D2-3gr10, D4-10g and D4-10gSS, which were described and applied in Chapter 3. With regard to the real-world data set, it is related with the parental recognition by hand's biometrics. We intend to investigate, by the consensus clustering techniques whether it is possible to find the parents of a child through the picture of his/her right hand. This problem has application, for example, to identify parents of people, lost at an early age, during natural calamities and wars.

Also, if we want to know who is the father (and mother) of a person, one cannot perform genetic testing to all the people, as it would be very expensive among many other restrictions. If the hands images constitute the database of the potentials, reducing the probable parents to a much smaller number, we are saving money. That's the idea of reducing the size of demand without the need to identify exactly who is the father or mother.

In order to carry these researches, the experiments are performed over the right hands images. Our database consists of right hand images of 187 people, whose are parents and children, and 3 hand images per person. 271 features per image were extracted using the algorithms available on, Bosphorus Hand Database [9] and also applied at the experiments in Chapter 6.

This hands images database was created to develop a parental biometric recognition system, and the unique constraint is the hand over a black background. All

the images were acquired through a normal mobile phone, in different situations of luminosity and proximity. These images initially saved as JPG images, were converted to map of bits with measures 382×525 bits, 588Kb, with color image resolution.

These experiences based on hands images were performed on families which are, fathers, mothers and children. Our goal is, to take a person (with 3 photos) and see if someone else corresponds to it, in terms of, father, mother or sibling. This will happen if they are placed together in the same cluster. Based on this data, we work on four different data sets, considering hands images of: 1) fathers and children (F); 2) mothers and children (M); 3) siblings (S) and 4) all the family, i.e., parents and children (P). Some of these images are illustrated in Figure 7.1.

Figure 7.1- Examples of hands images of six different people in our database.

First we focus on the results of the ARI values in Table 7.1 for the simulated data sets, in accordance with the data clusterings having clusters with different cardinalities even having close clusters as D1-4g and D4-10gSS, or overlapping clusters as D4-10g. In general the individual hierarchical clusterings show approximately the same performance and a good performance, as well as the consensus clustering techniques.

Now regarding the data set D3-2g, in which it has a data structure into 3 clusters all having the same cardinality. Also, 2 of them are close to each other and are equally compacts. The other cluster is apart and less compact. The results show that the individual hierarchical clusterings present different performances between them, as also the consensus clustering techniques. Besides, TEC.4 outperforms the other techniques, presenting a great performance (ARI value equals to 1). Adding data noise to this data set (D3-gr10), this affects the performance of some individual clusterings and consequently can affect the performance of consensus techniques.

Thus there are situations in which the consensus clustering derived by the techniques presents worse performance than some of the individual clusterings. Usually it happens applying the traditional consensus clustering techniques.

Some techniques are more affected by the performance of all the individual clusterings than other techniques. One can say that the performance of the traditional techniques is in accordance with the performance of the most of individual clusterings. On the other hand the multi-objective technique seems to be influenced by the clustering with good performance and does not by the most of them, as for data sets D2-3g and D4-10g. Moreover it can outperform any one the individual clusterings as for data set D4-10gSS.

Noting in Table 7.2 for the real-world data sets, derived from the database containing hand's images of parents and their children, the ARI values of consensus clustering techniques, do not reveal great performance.

The higher ARI value 0.3032, is obtained by TEC.2, for database having 3 hand's image of each father and 3 hand's image of his child. This ARI value reveals that the agreement between the clustering obtained by TEC.2 and the true clustering is

far to be perfect. It's clear that the TEC.2 produces a clustering in which haven't on the same cluster only the biometrics of the 3 hand images of a father and 3 hand images of his child, for all fathers and children. By other hand, this ARI value suggests that, it's possible that, for some father, the biometrics of at least one of his hand image and the biometrics of at least one hand image of his child are in the same cluster. Otherwise the ARI could be closer to 0.

The consensus clustering is obtained by the individual hierarchical clustering algorithms and these are formed by distances between clusters and elements, as was discussed in Chapter 2. So the impossibility of putting on the same cluster biometrics of 6 hand images (3 of a father and 3 of his child) means that at least some of these 6 hand's biometrics are not so close from all the others or from some other. So, we try to know by hands biometrics, how much far is a child from his/her parents. This is another analysis of this framework.

The procedure is, we get a person's photo and we calculate the distances to all other photos (by the biometrics). Having, each person 3 hand images, calculating the distances between each two person, we have 9 distances. Our statistic is the distance between each two people as the minimum of these 9 distances. Analysing the distribution of these distances for all people, allows us to verify, for instance, if "A" has his/her father, mother or sibling among 10% of the closest people. According to the probability of a child have his/her father, mother or sibling among 10% of the closest people. If this probability is for instance 95%, then the search for the parent of a child can be reduced for 10% of the closest people in the database.

In respect to the distances between people in the database, we search to fulfil the sentence: "Running the hand's images of a person on the database, where it is M, there is the probability P of M being identified among p of the closest people." We consider, M= {father, mother, sibling, at least one of these familiar} and p= {10%, 25%, 50%}. The probabilities P are in Table 7.3. According to these probabilities, we can state that, running the hand's images of a person by the database where there are the father, mother and a sibling, there is 95% of probability of at least one of these familiar be in the half of those closest people. This does not allow the identification of one family member but can restrict the search space for a half for instance, in a genetic test.

Table 7.1: For each simulated data set, the ARI values of the, **A**- individual clusterings; **B**- consensus clustering techniques.

A

Data set	Algorithm	ARI
D1-4g	SL	0.8143
	CL	0.9823
	AL	0.9823
	W	0.9823
D2-3g	SL	0.5584
	CL	0.4448
	AL	0.5584
	W	1
D2-3gr10	SL	0.3500
	CL	0.7937
	AL	0.3500
	W	0.7937
D4-10g	SL	0.7681
	CL	0.9518
	AL	0.9402
	W	0.9402
D4-10gSS	SL	0.9945
	CL	0.9946
	AL	0.9946
	W	0.9946

B

Data set	Technique	ARI
D1-4g	TEC.1	0.9823
	TEC.2	0.9823
	TEC.3	0.9823
	TEC.4	0.9823
D2-3g	TEC.1	0.5584
	TEC.2	0.5681
	TEC.3	0.5681
	TEC.4	**1**
D2-3gr10	TEC.1	0.7274
	TEC.2	0.7274
	TEC.3	0.7937
	TEC.4	0.7937
D4-10g	TEC.1	0.9402
	TEC.2	0.9377
	TEC.3	0.9402
	TEC.4	**0.9518**
D4-10gSS	TEC.1	0.9946
	TEC.2	0.9946
	TEC.3	0.9946
	TEC.4	**1**

Table 7.2: ARI values of the consensus clustering according to the database and the consensus clustering technique. Being **F**: fathers and children; **M**: mothers and children; **S**: siblings and **P**: parents and children.

Database	Technique	ARI
F	TEC.1	0.1571
	TEC.2	**0.3032**
	TEC.3	0.2280
	TEC.4	0.2463
M	TEC.1	0.2030
	TEC.2	**0.2901**
	TEC.3	0.2460
	TEC.4	0.2299
S	TEC.1	0.2711
	TEC.2	0.2875
	TEC.3	0.2414
	TEC.4	**0.2915**
P	TEC.1	0.1283
	TEC.2	0.2165
	TEC.3	0.1762
	TEC.4	0.1883

Table 7.3: The entries are the probability of M be among p of the closest people of a child.

M\p	**10%**	**25%**	**50%**
Father	29,3%	53,3%	79%
Mother	40%	57,5%	78%
Sibling	52,2%	76%	94%
A familiar	64,2%	85%	**95%**

7.5 Conclusions

In this Chapter we focused on the problem of to find the best consensus clustering. We analysed some of the approaches of consensus clustering most referred in literature. Such as, the traditional consensus clustering techniques addressed in Chapter 2, with different mechanisms to achieve the consensus clustering and a multi-objective consensus clustering technique. We proposed to analyse the performance of these consensus clustering techniques by matching the consensus obtained and the known clustering using ARI.

The base clusterings were obtained by the application of traditional hierarchical clustering algorithms, studied in Chapter 2.

We discussed these approaches by a comparative study considering a set of experiments using synthetic and real world data sets. For the simulated data sets, in most of the cases the multi-objective technique, MOCLE proved to outperform other techniques even in situations like noise introduction and clusters with different homogeneity or overlapping. Also, unlike other techniques it is less susceptible to the existence of individual clusterings which have poor quality. Moreover, it showed that it can capture the performance of the best base clusterings and still outperforms them.

Regarding the real data set, it is based on the hand's biometrics in context of parental recognition. With this data set we intended to investigate the possibility of parental recognition by the biometrics extracted from the hands images and applying the consensus clustering algorithms. The correctly identification of a child and her/his parents, could be a potential business in which a website says if A is B's son using photographs of hands, by the economic value of this technology.

Several researches have been developed on the area of personal recognition by hand's biometrics. Different systems have emerged, looking generally to get accuracy through the great personal recognition rate or others measurements. Moreover in the Chapter 6, by the hierarchical algorithms and the SEP/COP approach, we achieved 100% of recognition for some considerable hand set samples. The ability

to identify a person by his hand image can be helpful for instance for parental relationship identification. There are situations where it's necessary to identify whether a person is another person's child, for example, in the case of children that went missing. Although one can use a genetic test to identify the parenting of a child, the hand photography is fast, cheap, no need for a technical and can be used remotely to query an online database.

Regarding the application of consensus clustering techniques to these real data sets based on parents and children hands images, each of these images provides 271 biometrics. Intending to identify a person's father or mother by consensus clustering techniques, all the techniques presented approximately the same performance and it is not a good performance. This means that by the ARI values the parents and their children are not close enough to be placed in the same cluster. On the other hand, the ARI values also allow to conclude that consensus clustering and the real clustering are not in total disagreement, i.e., there is some proximity between parents and their children. So in another analysis, obtaining the distances between all the people in the database enabled us to reach a conclusion. Taking into account that, for someone having in the database his/her father, mother and a sibling, there is a great probability of at least one of them be in 50% of the closest people. This is a good result, although it doesn't identify the child's parents, instead, allows to reduce the domain of research substantially.

As final remarks we must refer that the most of the hands images were taken by a mobile phone at different conditions, such as luminosity, in which may be deficit. We believe that, being the collection of the images made by a scanner, the results by the consensus clustering techniques possible can be better. Thus, we consider this as future work.

Chapter 8

Conclusions and future work

Consensus clustering aims to combine multiple clusterings obtained from the same data set. It has revealed to be a better alternative than using a single clustering. Several consensus clustering techniques emerged in literature, each one with a specific way to combine the clusterings. So, different techniques applied to the same set of individual clusterings can provide different solutions. Moreover, these techniques always provide a consensus clustering even in situations where it might not have a consensus solution. These difficulties concerning to consensus clustering techniques (as discussed in Chapter 4), constitute the first part of our researches.

This thesis is composed by two parts. In the first part we proposed two goals. They are, to find conditions for the existence of consensus clustering and to find a new way to evaluate the consensus clustering.

For this, a set of experimental procedures was carried out considering, simulated data sets with some particular data structure into clusters and real data sets from the UCI Machine Learning Repository.

The consensus clustering techniques were applied to sets of base clusterings being the clusterings provided by the hierarchical clustering algorithms, namely, Single Linkage, Complete Linkage, Average Linkage and Ward with the Euclidian metric. The consensus clustering techniques applied were Voting K-means, based in voting mechanisms, EAC, based on co-association matrix and a consensus proposed by Strehl and Ghosh, based in hyper graphs and Mutual Information.

Proposing to give solution to our goals, we searched profiles of hierarchical clustering algorithms in terms of their variabilities and from these, we analysed the implication on the consensus clustering.

Our results showed that, by applying the technique based on hyper graphs and Mutual Information to the base clusterings with clusterings having great variability between them, it leads to a consensus clustering with quality.

This result allows to define a sufficient condition for the existence of consensus clustering, as well as, a new strategy to evaluate that.

The sufficient condition is defined by certain properties of the base clusterings. It is considering base clusterings, where the clusterings are provided by a hierarchical clustering algorithm and having great variability between them.

The new strategy to evaluate consensus clustering consists in measuring the variability of each set of base clusterings where the clusterings are provided by a hierarchical clustering algorithm. Then, considering the set of base clusterings, having the clusterings great variability between them, leads to the best consensus clustering.

Furthermore, the analysis of hierarchical clustering variability led to the study of a new property of hierarchical clustering algorithms which is described in here. Applying an algorithm better suited to a data set with certain characteristics of clusters, this algorithm presents small variability.

By the results above we can conclude that the consensus clustering obtained by the technique based in hyper graphs and Mutual Information may present a great performance under some conditions. These conditions are:

1. Considering a data set with a cluster structure and a hierarchical clustering algorithm, less suited to this data set;

2. Applying this hierarchical clustering algorithm, to data samples of this data set;

3. The resulting clusterings are the base clusterings for such consensus technique.

In the second part of this thesis, we proceeded to the applications of hierarchical clustering algorithms and consensus clustering techniques to the real-world data sets. The data sets derived from the hand's biometrics.

First, it was applied the usual hierarchical clustering algorithms and a different approach in literature to several data sets. This approach consists in a different post-processing on the hierarchy, SEP/COP.

These researches allowed us to find the SEP/COP algorithms outperform the usual hierarchical clustering algorithms and also outperform the results in literature. Namely considering databases with hands images of 50 people, it was achieved 100% of recognition. And for a database with hands images of 100 people, the recognition rate achieved was 99.16%.

Secondly, we proposed to investigate the relevance of consensus clustering techniques on data sets including the one based on parental recognition of people. Regarding this data set, first it was created the database with the hands images of parents and their children. To the biometrics of these hands images were applied the usual hierarchical clustering algorithms, which are the base clusterings for obtaining the consensus clustering. The consensus clustering techniques applied were the traditional ones as applied in the first part of this thesis, and the multi-objective MOCLE.

According to the results, despite no technique has presented a great performance, we discovered that the search for a person's parents can be restricted to

half of the database of the “closest” people with 95% of probability. This was done by calculating the distances between the biometrics of all hands, ordering it. This research contributes to an innovative work in applications involving the parental recognition by the hand’s biometry in which, consensus clustering algorithms managed to get a good advance at researches on this issue.

The perspectives to develop in future work consist in:

- Applying other clustering algorithms to the data sets with the biometrics of parents and children. Being the resulting clusterings the base clusterings of consensus clustering techniques. So, we can include here other ways to construct the base clusterings and also use other consensus clustering techniques;
- The extraction of other hand’s biometrics to the parental recognition issue;
- Other recognitions, for instance: 1- if the hand’s biometrics change over the time i.e., change with the person’s age; 2- if there are significant differences between the hand's biometrics of people from different races; 3- whether it is possible to identify risks of diseases known to have some degree of hereditary determination as diabetes and certain cancers; 4- there is also the possibility of behavioural characteristics be related to the hand geometry
- The construction of a new database of hand’s images of parents and children. These images should be collected by a scanner (in this work it was done by a digital camera). We intend to explore this new database, considering all the situations referred above. With this database with more quality, some of consensus clustering techniques analysed considering the framework studied in this thesis, can identify the parenting;
- Besides the personal recognition by hand’s biometrics, other real data sets can be used in consensus clustering analysis for instance, related to the renewable energies issues.

Bibliography

[1] Adan, M., A. Adan, A. Vasquez and R. Torres (2008). Biometric verification/identification based on hands natural layout, Image and Vision Computing, 26(4), pp. 451-465.

[2] Al-Razgan, M. and C. Domeniconi (2006). Weighted Clustering Ensembles, in Proceedings of the SIAM International Conference on Data Mining, Bethesda, Maryland, pp. 20-22.

[3] Albuquerque, M. A. (2005). Estabilidade em análise de agrupamento, dissertação apresentada a Universidade Federal Rural de Pernambuco para obtenção do título de Mestre em Biometria, Área de Concentração: Modelagem e Planejamento de experimentação.

[4] Anderberg, M. R. (1973). Cluster Analysis for Aplications: Probability and Mathematical Statistics, vol. 19, edit. Academic Press, pp. 142-148.

[5] Arbelaitz, O., I. Gurrutxaga, A. J. Lojo, J. Muguerza and I. Perona (2011). SAHN with SEP/COP and SPADE, to build a general web navigation adaptation system using server log information, Advances in Artificial Intelligence, Lecture Notes in Computer Science, 7023, Springer, pp. 413-422.

[6] Ayad, H. and M. Kamel (2003). Finding natural clusters using multi-clusterer combiner based on shared nearest neighbors, in Multiple Classifier Systems, Fourth International Workshop, MCS, LNCS 2709, UK, pp. 166-175.

[7] Bača, M., P. Grd and T. Fotak (2012). Basic Principles and Trends in Hand Geometry and Hand Shape Biometrics, New Trends and Development in Biometrics (Jucheng Yang, Shanjuan Xie, ed), InTech.

[8] Oliveira Faria, A. (2008). Biometria: Reconhecimento Facial Livre, Linha de Código.

[9] Bosphorus Hand Database. <http://bosphorus.ee.boun.edu.tr/hand/home.aspx>.

[10] Bulatov, Y., S. Jambawalikar, P. Kumar and S. Sethia (2004). Hand recognition using geometric classifiers, in proceedings of 1st ICBA, Hong Kong.

[11] Cardoso, M. G. M. S., K. Faceli and A. C. P. L. F. Carvalho (2010). Evaluation of Clustering Results: the trade-off Bias-Variability, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 201-208.

[12] Central source of information on biometrics-related activities of the Federal government (2006). <http://www.biometrics.gov/Documents/FaceRec.pdf>.

[13] Corne, D., N. Jerram, J. Knowles and M. Oates (2001). PESA-II: Region-based Selection in Evolutionary Multiobjetive Optimization, in Proc. of the Genetic and Evolutionary Computation Conference (GECCO-2001), San Francisco, California, USA, Morgan Kaufmann, pp. 283-290.

[14] Cover, T. M. and J. A. Thomas (2006). Elements of Information Theory, 2nd edn., Wiley Series in Telecommunications and Signal Processing (Wiley-Interscience).

[15] Richmond, Z. (2012). Hand Geometry, CPSC 601 Lecture Week 5.

[16] De Santos Sierra, A., C. Sanchez-Avila, G. Bailador del, Poz, and J. Guerra-Casanova (2011). Unconstrained and Contactless Hand Geometry Biometrics. Sensors, 11, pp. 10143-10164.

[17] Deb, K., A. Pratap, S. Agarwal and T. Meyrivan (2002). A Fast and Estilist Multiobjective Genetic Algorithm: NSGA-II, IEE Transion on Evolutionary Computation, 6(2), pp. 182-197.

[18] Dimitriadou, E., A. Weingessel and K. Hornik, (2001). Votingmerging: An ensemble method for clustering, in Proc. Int. Conf. on Artificial Neural Networks, Vienna, pp. 217-224.

[19] Ding, Y., D. Zhuang and K. Wang (2005). A study of hand vein recognition method, in Proceedings of 2005 IEEE International Conference Mechatronics and Automation, Canada, pp. 2106-2110.

[20] Domeniconi, C. and M. Al-Razgan (2009). Weigthed cluster ensembles: methods and analysis, ACM Trans. Knowl. Discov. Data 2(4), pp. 1-40.

[21] Duarte, F., J. Duarte, A. Fred and M. Rodrigues (2011). Cluster Ensemble Selection Using Average Cluster Consistency, Knowledge Discovery, Knowledge Engineering and Knowledge Management. Communications in Computer and Information Science, vol. 128, pp. 133-148.

[22] Dutagaci, H. and B. Sankur (2008). Comparative analysis of global hand appearance-based person recognition, Journal of Electronic Imaging 17(1), 011018.

[23] Ernie, G. N. (1977). Palm print identification, US patent 4032889 A.

[24] Ernst, R. H. (1971). Hand ID System, US patent 3576537 A.

[25] Faceli, K., A. Carvalho and M. de Souto (2009). Multi-objective clustering ensemble for gene expression data analysis, Neurocomputing 72, pp. 2753-2774.

[26] Faceli, K. (2007). Um framework para análise de agrupamento baseada na combinação multi-objetivo de algoritmos de agrupamento, in Tese apresentada ao ICMC-USP para obtenção do título de Doutor em Ciências de Computação e Matemática Computacional. USP.

[27] Ferrer, M. A., J. Fabregas, M. Faundez, J. B. Alonso and C. M. Travieso (2009). Hand Geometry Identification System Performance, In Proceedings of the 43rd Annual 2009 International Carnahan Conference on Security Technology, Switzerland, pp. 167-171.

[28] Fern, X. and C. Brodley (2004). Solving cluster ensemble problems by bipartite graph partitioning, on Proc. Of International Conference on Machine Learning, 36.

[29] Ferreira, L. and D. Hitchcock (2009). A Comparison of Hierarchical Methods for Clustering Functional Data, Communications in Statistics - Simulation and Computation, vol. 38, Issue 9, pp. 1925-1949.

[30] Fotak, T., P. Koruga and M. Baca (2012). Trends in hand geometry biometrics, Central European Conference on Information and Intelligent Systems, pp. 319-493.

[31] Fred, A. (2001). Finding consistent clusters in data partitions, in J. Kittler and F. Roli, editors, Multiple Classifier Systems, volume LNCS 2096. Springer, pp. 309-318.

[32] Fred, A. (2009). From Single Clustering to Ensemble Methods, Unsupervised Learning, Instituto de Telecomunicações, Instituto Superior Técnico da Universidade Técnica de Lisboa.

[33] Fred A. and A. Jain (2005). Combining Multiple Clusterings Using Evidence Accumulation, IEEE Trans Pattern Analysis and Machine Intelligence 27(6), pp. 835-850.

[34] Fred, A. and A. Lourenço (2008). Cluster Ensemble Methods: from Single Clusterings to Combined Solutions, Chapter in Supervised and Unsupervised Ensemble Methods and their Applications, Oleg Okun and Giorgio Ventini, Springer.

[35] Fridlyand, J. and S. Dudoit (2001). Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method, in Technical Report 600, Statistics Department, UC Berkeley.

[36] Futuristic Technology in Use at Olympic Games; Recognition Systems' Biometric Hand Geometry Product Used in High Security Venues in Atlanta; To Include the Olympic Residence of Both the Men's and Women's U.S. Basketball Teams. Business Wire, 1996.

[37] Gurrutxaga, I., I. Albisua, O. Arbelaitz, J. Martin, J. Muguerza, J. Perez and I. Perona (2010). SEP/COP: An efficient method to find the best partition in the hierarchical clustering based on a new cluster validity index, Pattern Recognition, 43, pp. 3364-3373.

[38] Hadjitodorov, S. T., L. I. Kuncheva and L. P. Todorova (2006). Experimental Comparison of Cluster Ensemble Methods, in Information Fusion, 9th International Conference on, pp. 1-7.

[39] Hadjitodorov, S. T., L. I. Kuncheva and L. P. Todorova (2006). Moderate diversity for better cluster ensembles, Information Fusion, vol. 7(3), pp. 264-275.

[40] Handl J. and J. Knowles (2004). Multiobjective clustering with automatic determination of the number of clusters, Technical Report TR-COMPSYSBIO-2004-02, MIST, Manchester.

[41] Handl, J., J. Knowles and D. B. Kell (2005). Computational cluster validation in post-genomic data analysis, Data and text mining, vol. 21(15), pp. 3201-3212.

[42] Hubert L. and P. Arabie (1985). Comparing Partitions. Journal of Classification 2, pp. 193-218.

[43] I-Cheng Y. (2008). Department of Information Management , Chung-Hua University, Hsin Chu, Taiwan 30067, R.O.C.

[44] Im, S. K., H. S. Choi and S. W. Kim (2003). A direction-based vascular pattern extraction algorithm for hand vascular pattern verification, ETRI J., 25, pp. 101-108.

[45] Jain, A.K., R. Bolle and S. Pankanti (1999). Biometrics: Personal Identification in Networked Society, Kluwer Academic Publications. ISBN 978-0-7923-8345-1.

[46] Jain, A. K. and R. C. Dubes (1988). Algorithms for clustering data, Ed. Prentice Hall, Inc.

[47] Jain, A. K. and N. Duta (1999). Deformable matching of hand shapes for verification, Proceedings of International Conference on Image Processing, Kobe, Japan, pp. 857-861.

[48] Jain, A. K., A. Ross and S. Pankansi (1999). A Prototype hand geometry-based verification system, In Proceedings of 2nd International Conference on Audio and Video-Based Biometric Person Authentication, Washington, DC, USA, pp. 166–171.

[49] Jain, A. K., A. Ross and S. Prabhakar (2004). An introduction to biometric recognition. IEEE Transactions on Circuits and Systems for video Technology, 14(1), pp. 4-20.

[50] Kanhangad, V., A. Kumar and D. Zhang (2010). Human Hand Identification with 3D Hand Pose Variations, In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, USA, pp. 17-21.

[51] Karypis, G., R. Aggarwal, V. Kumar and S. Shekhar (1997). Multilevel hypergraph patitioning: Applications in VLSI domain, in Proc. of the 34th annual Design and Automation Conference, pp. 526-529.

[52] Karypis, G. and V. Kumar (1998). A fast and high quality multilevel scheme for partitioning irregular graphs, SIAM Journal of Scientific Computing, Vol. 20, No. 1, pp. 359-392.

[53] Kerr, M. K. and G. A. Churchill (2001). Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments, PNAS July 31, vol. 98(16), pp. 8961-8965.

[54] Konukoglu, E., E. Yoruk, J. Darbon and B. Sankur (2006). Shape-based hand recognition, IEEE Trans. on Image Processing, 15(7), pp. 1803-1815.

[55] Krieger, A. M. and P. E. Green (1999). A cautionary note on using internal cross validation to select the number of clusters, PSYCHOMETRIKA, vol. 64(3), pp. 341-353.

[56] Kumar A. and C. Ravikanth (2009). Personal authentication using finger knuckle surface, IEEE Trans. Inf. Forensics Secur., 4, pp. 98-110.

[57] Kumar, Y. A., D. C. M. Wong, H. C. Shen and A. K. Jain (2003). Personal verification using palmprint and hand geometry biometric, Lecture Notes in Computer Science, 2688, pp. 668-678.

[58] Kumar A. and D. Zhang (2006). Personal recognition using hand shape and texture. IEEE Transactions on Image Processing, 15(8), pp. 2454-2461.

[59] Lange, T., M. L. Braun, V. Roth and J. M. Buhmann (2002). Stability-Based Model Selection, in Advances in Neural Information Processing Systems 15, pp. 617-624.

[60] Law, M. H. and A. K. Jain (2002). Cluster Validity by Bootstrapping Partitions, Department of Computer Science, Michigan State University, MSU-CSE-03-5.

[61] Law, M., A. Topchy and A. K. Jain (2004). Multiobjective data clustering, in Proc. Of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2, pp. 424-430.

[62] Levine, E. and E. Domany (2001). Resampling Method for Unsupervised Estimation of Cluster Validity, Neural Computation, vol. 13(11), pp. 2573-2593.

[63] Lee, E. C., H. C. Lee and K. R. Park (2009). Finger vein recognition using minutia-based alignment and local binary pattern-based feature extraction, Int. J. Imaging Syst. Technol, 19, pp. 179-186.

[64] Linden, R. (2009). Técnicas de Agrupamento, Revista de Sistemas de Informação da FSMA. N. 4, pp. 18-36.

[65] Liu, B. (2006). Web Data Mining - Exploring Hyperlinks, Contents and Usage Data, Springer, ISBN 3-540-37881-2.

[66] Liu, R., Y. Liu and Y. Li (2012). An Improved Method for Multi-objective Clustering Ensemble Algorithm, WCCI 2012 IEEE, pp. 10-15.

[67] Mangasarian, O. L., W. N. Street and W. H. Wolberg (1995). Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), July-August, pp. 570-577.

[68] Manning, C. D., P. Raghavan and H. Schütze (2008). Introduction to Information Retrieval, Cambridge University Press.

[69] Matos, H. (2011). Reconhecimento biométrico baseado na geometria da mão, Mestrado em Engenharia Eletrotécnica e Computadores, Major Automação.

[70] Mulyono D. and H. S. Jinn (2008). A study of finger vein biometric for personal identification, Biometrics and Security Technologies. In Proceedings of International Symposium on Biometrics and Security Technologies, Pakistan, pp. 1-8.

[71] Nakai, K. and M. Kanehisa (1991). Expert System for Predicting Protein Localization Sites in Gram-Negative Bacteria, PROTEINS: Structure, Function, and Genetics 11, pp. 95-110.

[72] Nelson, A.E., Y.M. Golightly, J.B. Renner, T.A. Schwartz, V.B. Kraus, C.G. Helmick and J.M. Jordan (2013). Differences in Multijoint Symptomatic Osteoarthritis Phenotypes by Race and Sex: The Johnston County

Osteoarthritis Project, ARTHRITIS & RHEUMATISM Vol. 65, No. 2, pp. 373–377.

[73] Öden, C., A. Erçil, V. Yildiz, H. Kirmizitas and B. Büke (2001). Hand Recognition Using Polynomials and geometric features, Lecture Notes in Computer Science, Volume 2091, Springer, pp. 336-341.

[74] Orwell, G. (1949). Nineteen Eighty-Four, Secker and Warburg, First Edition.

[75] Pascual, D., F. Pla and S. Sánchez (2010). Cluster validation using information stability measures, Pattern Recognition Letters, 31, pp. 454-461.

[76] Rahman, A., F. Anwar and S. Azad (2008). A Simple and Effective Technique for Human Verification with Hand Geometry, in Proceedings of the International Conference on Computer and Communication Engineering, Malaysia, pp. 1177-1180.

[77] Rand, W. (1971). Objective Criteria for the Evaluation of Clustering Methods, Journal of the American Statistical Association, vol. 66 (336), pp. 846–850.

[78] Roth, V., T. Lange, M. Braun and J. Buhmann (2002). A Resampling Approach to Cluster Validation, in Intl. Conf. on Computational Statistics.

[79] Sanchez-Reillo, R., C. Sanchez-Avila and A. Gonzales-Marcos (2000). Biometric identification through hand geometry measurements. IEEE Transactions of Patters Analysis and Machine Intelligence, 22, pp. 1171-1178.

[80] Sidlauskas, D. P. (1988). 3D hand profile identification apparatus, US patent 4736203.

[81] Silva, J., J. P. Marques de Sá and J. Jossinet (2000). Classification of Breast Tissue by Electrical Impedance Spectroscopy. Med & Bio Eng & Computing, 38, pp. 26-30.

[82] Sousa, L. and F. Sousa (2011). A procura da melhor partição em classificação hierárquica com recurso à abordagem SEP/COP. XVIII Jornadas de Classificação e Análise de Dados, Vila Real, 6-9 Abril.

[83] Sousa, L. and F. Sousa (2013). Análise de agrupamentos consensuais. XX Jornadas de Classificação e Análise de Dados, Universidade do Minho, 11-13 Abril.

[84] Sousa, L. and J. Gama (2014). The application of the hierarchical clustering algorithms for recognition using biometrics of the hand. International Journal of Advanced Engineering Research and Science, vol. 1, Issue-7, pp. 14-24.

[85] Sousa, L., J. Gama and K. Faceli (2014). Variability analysis of the hierarchical clustering algorithms and its implications on consensus clustering. TI Journals.

[86] Sousa, L., J. Gama and K. Faceli (2015). The consensos clustering analysis and an application to the parental recognition based on hand biometrics. IDA Journal.

[87] Strehl, A. and J. Ghosh (2002). Cluster Ensembles - A Knowledge Reuse Framework For Combining Partitionings, in Proc. Conference on Artificial Intelligence. Edmonton, pp. 93-98.

[88] Strehl, A. and J. Ghosh (2002). Cluster Ensembles - A Knowledge Reuse Framework For Combining Multiple Partitions, Journal of Machine Learning Research (3), pp. 583-617.

[89] Takashi, K. (1980). Identification apparatus, US patent 4206441 A.

[90] Terzi, E. (2009). Teaching Data Mining, CAS CS 565, Data Mining Fall.

[91] Tibshirani, R., G. Walther and T. Hastie (2001). Estimating the number of clusters in a data set via the gap statistic, Journal of the Royal Statistical Society: Series B, 63 (31), pp. 411-423.

[92] Topchy, A., A. K. Jain and W. Punch, (2003). Combining Multiple Weak Clusterings, in Proc. IEEE Int. Conf.on Data Mining, Melborne, FL, pp. 331-338.

[93] Topchy, A., A. K. Jain and W. Punch, (2004). A Mixture Model for Clustering Ensembles, in Proc. SIAM Conf. on Data Mining, pp. 379-390.

[94] UCI Machine Learning Repository. <https://archive.ics.uci.edu/ml/datasets.html>.

[95] Users Guide, SAS Institute. The Distance Procedure: Proximity Measures. Retrieved 2009-04-26.

[96] Vega-Pons, S., J. Correa-Morris and J. Ruiz-Shulcloper (2010). Weighted partition consensus via kernels, Patt. Recogn. vol. 43 (8), pp. 2712-2724.

[97] Vega-Pons, S. and J. Ruiz-Shulcloper (2011). A Survey of Clustering Ensemble Algorithms, Internation Journal of Pattern Recognition and Artificial Intelligence, vol. 25 (3), pp. 337-372.

[98] Wang, W. C., W. S. Chen and S. W. Shih (2009). Biometric Recognition by Fusing Palmprint and Hand-Geometry Based on Morphology, In Proceedings f the IEEE International Conference on Acoustics, Speech and Signal Proceeding, Taiwan, pp. 893-896.

[99] Warren, H. F., J. G. Anthony and H. J. Ian (1972). Personnel identification apparatus, US patent 3648240 A.

[100] Wu, X., D. Zhang and K. Wang (2003). Fisherpalms based palmprint recognition, Pattern Recognit., vol. 24, no. 15, pp. 2829-2838.

[101] Zhang, D., W. K. Kong, J, You and M. Wong (2003). On-line palmprint identification, IEEE Trans. Pattern Anal. Mach. Intell. Vol. 25, no. 9, pp. 1041-1050.