International Journal of Advanced Engineering Research and Science (IJAERS) *[Vol-4, Issue-6, Jun- 2017]*
https://dx.doi.org/10.22161/ijaers.4.6.14 *ISSN: 2349-6495(P) | 2456-1908(O)*

Variability analysis of the hierarchical clustering algoritms and its implication on consensus clustering

Lúcia Sousa

Superior School of Technology and Management of Polytechnic Institute of Viseu, Portugal

Abstract— *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 underlying structure of the data sets taking into account clustering criteria, properties in the data and specific way of data comparison. In the literature many clustering algorithms have been proposed having a common goal which is, given a set of objects, grouping similar objects in the same cluster and dissimilar objects in different clusters.*

Hierarchical clustering algorithms are of great importance in data analysis providing knowledge about the data structure. Due to the graphical representation of the resultant partitions, through a dendrogram, may give more information than the clustering obtained by non hierarchical clustering algorithms. The use of different clustering methods for the same data set, or the use of the same clustering method but with different initializations (different parameters), can produce different clustering. So several studies have been concerned with validate the resulting clustering analyzing them in terms of stability / variability, and also, there has been an increasing interest on the problem of determining a consensus clustering.

This work empirically analyzes the clustering variability delivered by hierarchical algorithms, and some consensus clustering techniques are also investigated. By the variability of hierarchical clustering, we select the most suitable consensus clustering technique existing in literature. Results on a range of synthetic and real data sets reveal significant differences of the variability of hierarchical clustering as well as different performances of the consensus clustering techniques.

Keywords* — *Data Mining, Cluster analysis, Consensus clustering, Hierarchical clustering algorithm, Validation indices.

I. INTRODUCTION

The clustering algorithms are much applied in Data Mining, and widely used in solving real problems from various fields such as Medicine, Psychology, Botany, Sociology, Biology, Archeology, Marketing, etc. [28]. They are unsupervised learning algorithms aiming to find a clustering of a given data set, such that, similar elements belong to 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 due to the visualization of the clustering through a dendogram. Different hierarchical clustering algorithms are proper for different shaped clusters, so may produce different clustering. Thus, putting up the problem of choosing one of these clustering (which is not a trivial task), or determines a clustering that represents the consensus among these clustering.

The difficult task of choose one clustering can be based on evaluating the clustering quality. The analysis of compactness and separation of clusters not always find the real clusters [3]. Furthermore, property as variability or stability, enable us to meet more stable solutions and infer about clustering quality. On the other hand, many works have sought combine the different clustering obtained by different algorithms and still get the best data clustering, namely, a consensus clustering, which a better clustering often means a more stable, more robust and more consistent clustering.

Several approaches to produce consensus clustering have been proposed and carried out in various ways which may lead to different consensus clustering for the same base clusterings set. Furthermore, some works to evaluate/select the best consensus clustering have been proposed in literature. As, in [14] is proposed a diversity measure of the base clusterings and its relation to the consensus clustering quality. Also, in [5] the authors propose measures to select the best consensus, based on consistency between the base clusterings and the consensus clustering. In this work, in order to select the best consensus clustering, we propose to analyze the variance of the base clusterings and its relation to the consensus quality.

The quality of a consensus clustering algorithm is measured by the match between the clustering obtained and the known truthful clustering of the data set. From some matching indices suggested in the literature, we apply the Adjusted Rand index and Normalized Mutual Information, because they are, perhaps, the most popular ones, quantifying the proportion of pairs in agreement of two clustering informing if two clustering are independent from one another. The variability of the base clusterings set is obtained by the match between two by two clustering and is calculated by the standard deviation of Adjusted Rand index as in [3].

The base clusterings set is obtained by hierarchical clustering algorithms, namely, Single-Linkage, Complete-Linkage, Average-Linkage and the Ward method. To these clustering three consensus clustering techniques much reported in the literature are applied. One based on voting mechanisms, other is based on co-association matrix (EAC) and another of them is based on Mutual Information and hyper graphs. Our investigation is considering artificial and real data sets, being the artificial data, with different characteristics, in terms of number of clusters, cardinality, cohesion and separability, furthermore, for the real data sets also are considered different dimensionalities.

The remaining of this paper is organized as follows. In Section 2, we introduce the related work, in which, we address some known characteristics of hierarchical clustering algorithms, the consensus clustering techniques of interest for this work, as well as validation indices used for the analysis and different ways to select/validate the consensus clustering. In Section 3, we focus some existing alternatives to analyze the clustering variability, and also is described the methodology used to quantify the clustering variability and by this how to achieve the consensus clustering. In Section 4, we perform a set of experiments in order to analyze the variability of the hierarchical algorithms and the relation between the clustering variability and the performance of the consensus clustering techniques. In Section 5, conclusions are provided.

II. RELATED WORK

In this section we outline some related subjects with this work, such as, the differences between hierarchical clustering algorithms, the main approaches of consensus clustering, as well as the clustering validation issue. In a latter context, are discussed works concerned about the selection of the consensus clustering, by the application of clustering algorithms and validation indices.

A. Hierarchical clustering algorithms

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 [18].

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 Expectation-Maximization, but instead, they do not scale well and have, at least, time complexity of $O(n^2)$, where n is the number of elements [30], [6].

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, according to the distances between clusters, 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 obtaining clusters with the individual elements [30],[26]. 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. As the dendrogram usually contains more than one partition having different number of clusters, at our studies, we decide to fix the cut level of the dendrogram, i.e., fix the number of clusters according the data sets and their known structure.

Different hierarchical clustering algorithms differ on definition of distance between clusters henceforth may conduct to different resulting clusterings. The Single Linkage (SL) method compute the distance between two clusters by the minimal distance between all elements one of each cluster. For Complete Linkage (CL) method the distance between two clusters is the maximal distance between all elements one of each cluster. Considering Average Linkage (AL) method the distance between two clusters is the average distance between all pairs of elements, one in each cluster. The Ward’s method (W), also known by the method of minimum variance, differs from the above mentioned methods for not using distances between clusters to aggregate them. The objective of W is to look at the slightest deviation between the cluster centroid and the others elements of the cluster, i.e., looks at the smallest variance of the cluster. At each step, all the possibilities of adding two clusters are checked, and it’s chosen the one which causes the smallest increase of the sum of squares error, SSE, of the aggregate cluster.

Being, SSE=$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2$, k the number of clusters, y_{ij} the j^{th} element in the i^{th} cluster having centroid $\bar{y}_{i.}$ and n_i elements.

The distances between clusters are computed by distances between two elements, in which can be for instance, Euclidian, Mahhattan or Mahalanobis distance.

At this work we chose the Euclidian distance because, in ours preliminary experiments, this metric, was found 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 [2].

Having different definitions of distance between clusters the hierarchical clustering algorithms may produce different resultant partitions for the same data set. 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, SL can produce clusters disordered, elongated and little compacts [30]. 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 small 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 [30]. 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 [18]. 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 [30]. On the other hand, W, seeking to minimize the deviations between, cluster's elements and cluster's mean; it's an indication of homogeneity. The distance between two clusters is defined as the consequent increase in SSE if both clusters would join to form a single cluster. W algorithm, is attractive because it is based on a measure with strong statistical, and generate clusters, as well as CL, having a high internal consistency. Also has better performance than other hierarchical methods, especially, when the cluster's proportions are approximately equal [7]. Some principal characteristics of the SL, CL, AL and W algorithms are established in the Table 1.

Table1: Main properties of SL, CL, AL and W algorithms.

SL[28,39]	**CL[10,18,39]**	**AL[30]**	**W[1,2,7]**
Favors connectivity of clusters.	Favors compactness of clusters.	Clusters tend to spherical shapes.	Favors compactness of 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.		

B. Consensus clustering algorithms

As each hierarchical clustering algorithm has its own characteristics, the application of different clustering algorithms, may generate a wide variety of solutions, for a given data set. Faced with the existence of different clustering algorithms, initially, some authors were worried about searching for a particular algorithm which produces a given clustering configuration that best fits the data set, but, lately the investigation turned to the problem of how to combine the different clustering delivered by different algorithms. Several contributions to this problem have emerged in the literature, in which the combination of different clustering, aims to obtain a "better" data clustering, which represents the consensus among these clustering [10].

The various techniques in processing consensus clustering consist of two principal steps, one is Generation, which defines how to produce the set of individuals clustering, and the other is Consensus Function, describing how to combine them to find the consensus clustering. Thus, different ways to obtain and combine clustering lead to different consensus clustering techniques. Furthermore, each technique considers that certain properties should be fulfilled by the consensus clustering. These properties can be, i) Stability- Lower sensibility to noise or outliers, ii) Consistency- Similar to all the individuals clustering, iii) Robustness- Better performance than the individuals clustering and iv) Novelty- A clustering different from the individuals [11].

In the Generation step, there are no constrains about how the clustering must be obtained. Therefore, different clustering algorithms or the same algorithm with different parameters initialization can be applied. A common idea in 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 [14]. At the second step, the Consensus Function focuses the methodology of combining these individuals clustering to obtaining the consensus clustering. The Consensus Function is the main step for any consensus clustering algorithm and can be based, for instance, on Voting, Co-association Matrix, Graph and Hyper graph Partitioning, Information Theory, Finite Mixture Models, Genetic Algorithms. Moreover, some consensus functions are based on more than one of these approaches [11].

From several important contributions in the consensus clustering framework, one should note the works of, Fred [8], Fred and Jain [9] and Strehl and Ghosh [33-34], which are the pioneers in traditional consensus clustering approaches and are perhaps, the most referred in the literature. Due to that, we chose these consensus clustering techniques for our studies.

In [8], the Consensus Function is based on Voting and Co-association Matrix. The objective is to find consistent and robust consensus clustering. The individuals clustering are delivered using the K-means algorithm. With the data clustering obtained, pairs of elements are voted to be in the same cluster on consensus clustering every time they belong to the same cluster in the different clustering. The number of times that pair of elements is 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 higher than 0.5 (the threshold pre-defined). That means that pairs of elements are in the same cluster in more than 50% of individuals clustering.

The EAC (Evidence Accumulation Clustering), consists of a modification of [8] where the co-association matrix is represented as a graph [9]. 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. Wherein for each level is delivered a clustering with k clusters, and one range with the highest value is selected as the consensus clustering [11].

In order to build robust consensus clustering, in [33-34], 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 on by applying Mutual Information and representation on hyper graphs. The Mutual Information, concept from Information Theory [4] is used to measure the shared information between pairs of clustering. The consensus clustering is a clustering that shares most information with all possible clustering. The objective of finding a clustering that maximizes the Mutual Information, by an exhaustive search of pairs of clustering, raises computational problems. To solve this problem, three algorithms based on a hyper graph representation and partitioning algorithms are proposed, CSPA - Cluster-based Similarity Partitioning Algorithm, HGPA – Hyper Graph Partitioning Algorithm and MCLA - Meta-Clustering Algorithm. The result of each of these algorithms is a consensus clustering. The three algorithms start from representing the individuals clustering as a hyper graph, where each clustering is represented by a hyper edge. The CSPA algorithm constructs a co-association matrix where its values are weights associated to each two elements (nodes), corresponding on hyper graph representation, to the edge between the elements. After that, it’s applied the graph partitioning algorithm METIS that reduces the size of the graph by collapsing the vertices and edges, and after getting a partition from the smaller graph, the METIS then uncoarsen it to construct a partition for the original graph [20]. The greater the weight of the edge, the greater is the similarity between elements. Thus, on the first phase of METIS, this is the criterion used to join the common vertices, edge with the highest weight. The partition obtained by the smaller graph, is through an algorithm based on similarities. The HGPA algorithm applies also a partitioning algorithm, HMETIS, corresponding to hyper graphs [21]. 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 elements grouped in respective clusters. In hyper graph representation the clusters are nodes and the edges between two nodes have weight which is the similarity between the clusters. By the partitioning algorithm METIS, one obtains clusters called meta-clusters, and is calculated the times that each element appears in a meta-cluster. Being each element assigned to the meta-cluster to which appears more often [11]. Now, from these consensus clustering (associated to the three algorithms) is possible to search for final consensus clustering, the one which maximizes the shared Mutual Information. These authors, unlike the previous ones, use different algorithms to obtain the individuals clustering, and also pre define the desired number of clusters in the consensus clustering.

C. Clustering validation indices

Cluster validity can provide a quantitative answer, through validation indices, for the need of validate the output of a clustering algorithm. A validity index can be seen as a factor which assesses the goodness of a clustering [25]. The validation indices are applied according to the criteria employed which can be classified as external or internal criterion. Regarding to the external criteria a clustering is evaluated by the knowledge of a truly data clustering and according this criteria the usual indices applied are the, for instance, the Adjusted Rand [16] and Normalized Mutual Information [33-34].

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

The ARI is based on agreements and disagreements of pairs of elements of two clustering and are computed by the equation (1). Where, U and V are two different clustering of the data set, n is the number of elements, the clustering U has R clusters, and the clustering V has C clusters, 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.

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

In Information Theory, the Normalized Mutual Information (NMI) is a symmetric measure to quantify the statistical information shared between two distributions [33-34].

Considering the two clustering U and V and the same descriptions of the terms of the ARI's expression, as above, the NMI is given by the equation (2).

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

ARI and NMI can take values in the interval [0,1]. The value equal to 1, means perfect agreement between the two clustering unlike the values close to 0 (even negative values for ARI) indicating total disagreement.

D. The combination of the clustering algorithms, consensus clustering algorithms and clustering validation indices

Faced with the existence of different techniques to build the consensus clustering, some works have been worried about the problem of validate the resulting consensus clustering.

We describe below some experiments proposed to compare the performance of different consensus clustering, taking into account some measure which identifies the base clusterings that lead to the best consensus clustering.

Let Z be a set of n data, let P=$\{C_1,\dots,C_K\}$ be a clustering of Z into K clusters. A base clusterings set P is as set of N clustering of Z, P=$\{P_1,\dots,P_N\}$. Let P^* be a consensus clustering and P^T be the true clustering of the data.

In [14], the authors propose four diversity measures for the base clusterings and the consensus clustering, based on ARI. The various 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, $Div_1(P,P^*)$, is defined as the average diversity between each clustering $P_i\in P$ and the consensus clustering, P^*. It can be seen in Equation (3), where $AR(P_i,P^*)$ is the ARI value between the pairs of data clustering P_i and P^*, and $1-AR(P_i,P^*)$ is the diversity of the individual clusterings. The second measure $Div_2(P,P^*)$ is defined as the standard deviation of the diversity of the individual clusterings (Equation (4)). The third and forth diversity measures, $Div_3(P,P^*)$ and $Div_4(P,P^*)$ are derived from the first and second ones, and can be seen in Equations (5) and (6), respectively. The accuracy of the consensus clustering, P^*, is calculated as $AR(\,P^T\,,P^*)$.

$$Div_1(P,P^*)=\frac{1}{N}\sum_{i=1}^{N}\left(1-AR(P_i,P^*)\right) \quad (3)$$

$$Div_2(P,P^*) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(1 - AR(P_i,P^*) - Div_1(P,P^*)\right)^2} \qquad (4)$$

$$Div_3(P,P^*) = \frac{1}{2}\left(1 - Div_1(P,P^*) + Div_2(P,P^*)\right) \qquad (5)$$

$$Div_4(P,P^*) = \frac{Div_2(P,P^*)}{Div_1(P,P^*)} \qquad (6)$$

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, and that one should select the base clusterings with median values of $Div_1(P,P^*)$ or $Div_3(P,P^*)$ to get the best consensus clustering.

In another work [13] the authors evaluate the accuracy of the consensus clustering using 24 different scenarios, each one describing the base clustering algorithms and the consensus function applied. The base clustering algorithms used are, K-means, SL, AL and also these algorithms considering sub samples of the data. 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 like in [14]. After performed a set of experiments comparing the different scenarios, they conclude that the best can be using base clusterings obtained by K-means algorithms and the consensus function in which interpret the consensus matrix of the base clusterings as data instead of similarity.

In [5] the authors propose a new measure, to select the best consensus clustering among a variety of them. This measure is based on a concept of average cluster consistency, $ACC(P,P^*)$, which measures the average similarity between each clustering P_i of the base clusterings and a consensus clustering P^*. The definitions of measures can be seen by Equations (7) and (8), where, $K_i \geq K^*$, being K_i and K^* the number of clusters of the clustering P_i and P^*, respectively, and $|Inters_{kj}|$ 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, $Ci(P^T, P^*)$ [8], which measures the quantity of data shared in matching clusters of the real clustering and the consensus clustering and it is defined by Equation (9), where K^T is the number of clusters of the true clustering.

$$ACC(P,P^*) = \frac{1}{N}\sum_{i=1}^{N} sim(P_i,P^*) \qquad (7)$$

$$sim(P_i,P^*) = \frac{1}{n}\sum_{j=1}^{K_i} \max_{1\leq k\leq K^*} |Inters_{kj}| \left(1 - \frac{|C_{K^*}|}{n}\right) \qquad (8)$$

$$Ci(P^T, P^*) = \frac{1}{n}\sum_{k=1}^{min\{K^*,K^T\}} |C_{K^*} \cap C_{K^T}| \qquad (9)$$

In the experiences, the base clusterings are obtained, among others algorithms, by K-means, SL, AL, CL, and also considering join clustering obtained by these algorithms. The number of clusters is randomly chosen between 10 and 30. The consensus clustering is obtained by the EAC technique and also by others two variants of the WEACS technique. This technique is an extension of the EAC, being the weighted co-association matrix and using sampling of the data. The accuracy of a consensus clustering is 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 $ACC(P,P^*)$ value.

III. CLUSTERING VARIABILITY/ STABILITY AND OUR WORK

A. *Clustering variability and stability*

Many authors for the purpose of validate clustering, analyze the stability / variability / diversity of the clustering obtained by data resampling. The different works differ on the following issues: i) The methodology for resampling data, as, bootstrap [22], [25] or cross-validation [23], [24], [35], [3], [32]; ii) Clustering algorithm applied to the samples, as, K-means and hierarchical [23], K-means and EM [3], K-means, EM and hierarchical [25], [32] or K-means, KNN and hierarchical [27]; iii) Validation criteria, as, internal [22-23] or external [15]; iv) Validation indices, as, Gap [24], Adjusted Rand [23,15,3] or based on Information Theory [3], [32].

As the interest of this paper is about the clustering algorithm variability, one can mention some work concerned with this, existing in the literature, as for instance, the work in [25], in which, the authors interpret an algorithm of clustering as a statistical estimator and examine the variability of this estimator. This variability can be described as follows.

Considering a data set with size n, *Y*, get *k* samples, by resampling, each one with the same size n, $Y^1, \ldots, Y^k$. To apply to each sample, a clustering algorithm, designated by

A, obtaining then, *k* clustering, $A(Y^1),\dots,A(Y^k)$. The variability, *V*, of the clustering algorithm *A* is obtained by Equation (10), where, d, measures the distance between two clusterings and can be done by any measure of similarity between clusterings, as the indices, ARI, Jaccard, Folkes & Malows and Hubert. The value of *V* low means that the clustering algorithm is stable.

$$V = \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j=i+1}^{k} d(A(Y^i), A(Y^j)) \qquad (10)$$

Another work in [3] analyzes the variability of a clustering by data resampling based on a weighted cross-validation procedure. From 20 weighted samples and the original sample moreover by a clustering algorithm as K-means, one gets clusterings for the original sample and for the weighted samples. It is measured the agreement between the clustering of the original sample and each one of the clustering of the weighted samples, by the Adjusted Rand index. Once having the 20 values of the Adjusted Rand index, its standard deviation is used to measure the clustering variability.

B. *Our work*

In this study, considering the hierarchical algorithms, we propose to evaluate the clustering variability by external criteria, and from this, the implications on the performance of three consensus clustering techniques.

The comparison between the clusterings obtained is made by ARI and the measure of the clustering's variability is the standard deviation of ARI [3]. From these clustering, it is applied the consensus clustering techniques referred, and to evaluate the accuracy of these techniques, are applied, ARI and NMI, which have very similar behavior.

Intending to analyze the clustering variability delivered by hierarchical algorithms, the first hypothesis under study is, whether the different processing forms of the hierarchical clustering, affects the respective variability.

Regarding to the other hypothesis about the consensus clustering, we perform some studies to analyze the performance of some consensus clustering techniques, taking account the variability of the hierarchical base clusterings set, therefore, the second hypothesis under study is whether the performance of the consensus techniques depends on the variability of the base clusterings set.

To test these hypotheses, a set of experiments are implemented.

IV. EXPERIMENTAL DESIGN

The following subsections, report the experiments in order to validate the hypotheses under study.

A. *Data sets*

In order to reach the variety of situations regarding to the data sets, different data sets simulated and real are considered. The differences are with respect to cardinality, number of cluster, the shape of the clusters, as, well separated clusters and quite close clusters and clusters with distinct densities. Also it is considered data sets with added noise and with overlapped clusters. A description for each data set is given below.

1. *Simulated data sets*

In Fig. 1 to Fig. 7 are represented the 2-dimensional simulated data sets used in our experiments and in the Table 2 are the details of those data. The data sets have random data (according to their partition into clusters) and Normal distribution. Some of them are data sets used by others papers. On some data sets, noises randomly uniformly distributed are added. There are seven data sets assigned, D1-4g, D2-3g, D2-3gr10 (data sets D2-3g, with 10% noise), D3-3g, D3-3gr10 (data sets D3-3g, with 10% noise), D4-10g [12] (data set having overlapped clusters) and D4-10gSS [12] (data set D4-10g, without overlapped clusters).

2. *Real data sets*

In the experiments we apply seven real data sets which are taken from the UCI Machine Learning Repository [19]. These data sets, besides different cardinalities, number of clusters and shape of the clusters, also have different dimensionality, wherein, some of them are used in medical studies. These data sets are described below and a summarized in the Table 3.

- Iris: Refer to types of iris flowers. The attributes are four, sepals length, sepals width, petals length and petals width. The clusters of iris plant are, Setosa, Versicolour and Virginica.
- Ecoli: The clusters describe protein localization sites in Gram-negative bacteria E.coli [31].
- 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 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 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, which, in one cluster are the patients that survived at least 5 years and the other cluster has the patients which 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 data base. The four attributes are:

Recency – months since last donation, Frequency - total number of donation, Monetary - total blood donated, and Time - months since first donation. The data are then divided into two clusters representing whether the donor donated blood in March 2007 (yes or no) [17].

- WDBC- Wisconsin Diagnostic Breast Cancer: Contains 30 variables computed from digitized images of aspirated fine needle of a breast mass, which describing the characteristics of a cell nuclei presents. The clusters are two, meaning the diagnosis, benign or malignant [29].
- Breast Tissue: Consists of measures of electrical impedance of tissue samples taken freshly from the breast. This data can be split into six clusters, Carcinoma, Fibro-adenoma, Mastopathy, Glandular, Connective and Adipose [36].

Table 2: 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 elements for cluster i, OC and AN means overlapped clusters and add noise, respectively. The data noise are generated by Uniform distribution U(a,b) where (a,b) is the support interval.

	C	**Ni**	**Source**	**OC**	**AN**
D1-4g	4	15×35×35×35	C1: $N((0.5,0) , (0.05,0.05))$, C2: $N((-1,4) , (0.2,0.2))$ C3: $N((2,0) , (0.2,0.2))$, C4: $N((2,3.5) , (0.2,0.2))$	No	No
D2-3g	3	3×50	C1: $N((-1,0) , (0.25,0.25))$, C2: $N((1.5,2.5) , (0.25,0.25))$ C3: $N((8.5,10) , (2.25,2.25))$		
D2-3gr10	3	50×56×59	C1: $N((-1,0) , (0.25,0.25))$, C2: $N((1.5,2.5) , (0.25,0.25))$, U(3,4) C3: $N((8.5,10) , (1.5,2.25))$, U(6,7)		Yes
D3-3g	3	3×100	C1: $N((-1,-1) , (0.5,0.5))$, C2: $N((2,2) , (0.7,0.7))$ C3: $N((-3,3) , (0.1,0.1))$		No
D3-3gr10	3	130×100×100	C1: $N((-1,-1) , (0.5,0.5))$, U(0,0.3) C2: $N((2,2) , (0.7,0.7))$, C3: $N((-3,3) , (0.1,0.1))$		Yes
D4-10g	10	25×5 50×5	Ci: $N(([0, 50], [0, 50]) , ([0.1,$ (i=1,..10.	Yes	No
D4-10gSS	10	25×5 50×5	Ci $N(([0, 50], [0, 50]) , ([0.1,$ 0 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 center points respectively [12].	No	

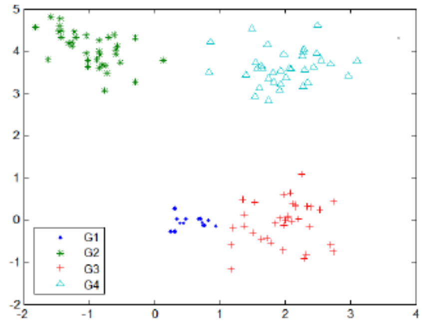

Fig. 1- Representation of data set D1-4g.

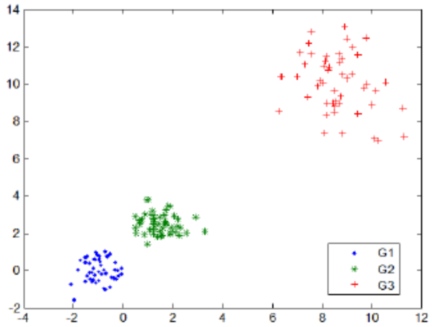

Fig. 2- Representation of data set D2-3g.

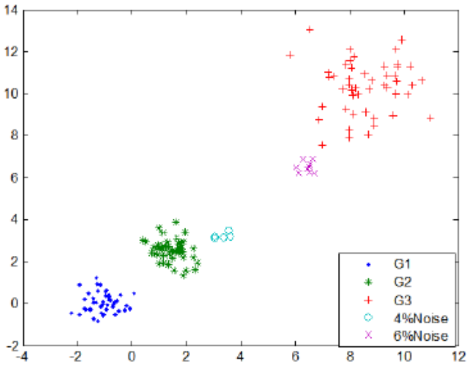

Fig. 3- Representation of data set D2-3gr10.

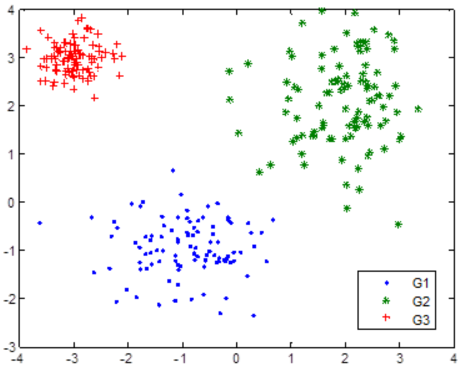

Fig. 4- Representation of data set D3-3g.

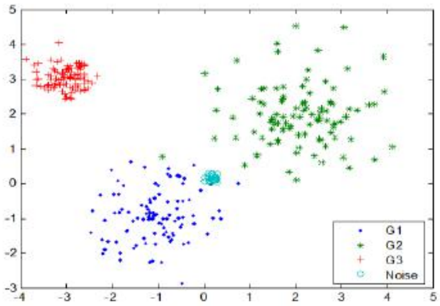

Fig. 5- Representation of data set D3-3gr10.

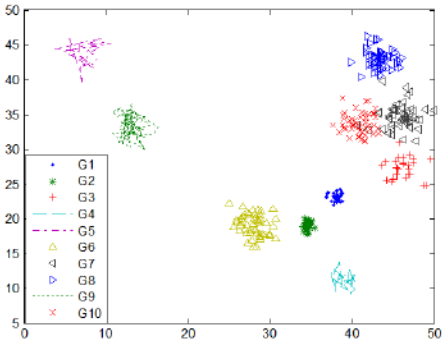

Fig. 6- Representation of data set D4-10g.

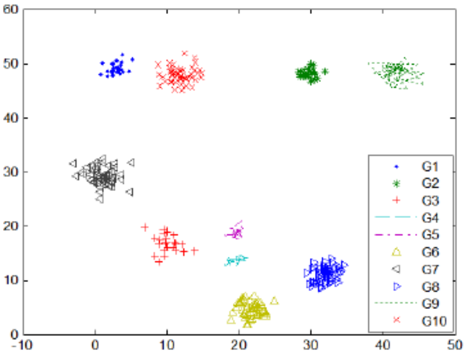

Fig. 7- Representation of data set D4-10gSS.

Table 3: Real data sets Summary. N is the number of data elements, C is the number of clusters and D is the dimensionality.

Name	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

B. *Generation of the base clusterings*

Intending to produce the base clusterings set, to each data set are applied the clustering algorithms, SL, CL, AL and W (with the Euclidean distance).

For each data set, it is considered data resampling without replacement, yielding 50 data samples of size (2⁄3)N, where N is the cardinality of the data set. For the real data sets, before the resample, first the data are normalized to mean 0 and standard deviation 1. Each clustering algorithm is applied to samples, obtaining the corresponding base clusterings set.

As the hierarchical algorithms produce a hierarchy of partitions, cutting the dendrogram in accordance with the number of pre-established clusters, results in a clustering. So, each base clusterings set delivered has the same number of clusters according to the known data partition.

To analyze the variability of the base clusterings set, the clustering are compared to each other only on the data shared by them. Taking account that to get the consensus clustering all the base clusterings must have the same data, it is added to each clustering the remained data, from the data set, that were not selected in the sample.

C. *Obtaining the consensus clustering*

For each base clusterings set, to generate the consensus clustering three consensus clustering techniques are applied, namely one based on Voting scheme [8] (TEC.1); *Evidence Accumulation Clustering* [9] (TEC.2) and other based on Mutual Information and Hypergraphs [33, 34] (TEC.3).

D. *Results and discussion*

1. *Variability of hierarchical clustering algorithms*

Given the data set and the clustering algorithm, from the 50 base clusterings obtained, it is calculated the ARI between them and consequently the measure of clustering variability which is the average ARI value. These results are stated in the Table 4.

In order to compare the variability of the hierarchical clustering algorithms, it is applied the hypothesis test (unilateral) of variances' equality, the F Snedecor test. Wherein, we can statistically conclude about the relation of the clustering algorithms variances. In the Table 5 are displayed these relations.

Table 4: Comparison of the hierarchical clustering techniques by the ARI average and the measure of variability, for each data set. The best relative results are highlighted.

	Data set	Algorithm	Avg	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 5: Relations of the hierarchical clustering's variances by the F Snedecor statistical test, for each data set.

Name	Relation
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

Analyzing the variability results in the tables 4 and 5, for almost all the data sets, the clustering algorithm which presents greater average ARI also presents the lowest variability, with exceptions, for the simulated data set, D4-10g and the real data set Blood.

Regarding the simulated and real data sets, W and AL present at almost all the data sets, the lowest variability, and at one of the cases, W achieves variability equal to 0 and average ARI equal to 1. By other hand, SL presents at almost all the data sets the greater variability with the exception of D2-3gr10, Iris and Ecoli data sets.

For some data sets, some clustering algorithms present equal and smaller variability than the remaining algorithms. For instance, for the data set Ecoli, SL and AL clustering algorithms and for data sets, Haberman's Survival and Breast Tissue, CL, AL and W clustering algorithms.

Observing the effect of data noise on variability, it is noted that for data sets D2-3gr10 and D3-3gr10, the CL clustering algorithm show the relatively most sensitivity to the noise. Regarding data sets D4-10g and D4-10gSS, all the clustering algorithms are affected by overlapping clusters.

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

Now, analyzing the graphic representation with the characteristics of the simulated data sets, and taking into account the differences between the hierarchical algorithms, as well as, the result of their variability, we can set the following statements.

- Considering the data set D1-4g, where 3 clusters (C2, C3 and C4) despite have the same cardinality and cohesion, they have greater variance regarding to the remaining cluster, so they are not compact and neither elongated (see Table 2 and Fig. 1). It is somehow expected that the SL and CL produces less stability, and is mainly due to the result of its higher variability in relation to AL and W.

- For data set D2-3g, having all clusters the same cardinalities, C1 and C2 have smaller variance than the remaining cluster, are then more compact, also smalls with spherical shape and close to each other (see Table 2 and Fig. 2). After that, is expected that CL and W produce more stable clustering, according to the lowest variability of these clustering in relation to SL and AL.
- With regard to data set D3-3g, where all the clusters have the same cardinalities and spherical shapes, 2 of them (C1 and C2) are less compact than the remaining one, also slightly apart and having larger diameters (see Table 2 and Fig. 4). It is expected that SL are less stable and moreover, presents a higher variability compared to the others clustering algorithms.
- Taking account the data set D4-10gSS (without overlapped clusters), wherein the clusters are different from each other, have different cardinalities, in general, they are compacts and some of them slightly separated (see Table 2 and Fig. 7), it is expected that SL clustering cope less stability, resulting in higher variability, with regard to the remaining clustering algorithms.
- Regarding to the data set D4-10g, having overlapped clusters (see Table 2 and Fig. 6), the variability values of all the clustering algorithms increase in relation of the corresponding data set without overlapped clusters.
- As CL clustering algorithm is more sensitive to outliers or noisy data, the variability values for data sets D2-3gr10 and D3-3gr10 (see Table 2 and Figs. 4, 6) are expected.

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

2. *Impact on consensus*

In order to compare the consensus clustering obtained by the three techniques with the known clustering of the data sets, the ARI and also the NMI are calculated. For each data set and each base clusterings derived by the hierarchical algorithms, the Table 6 contains the ARI and NMI values for each consensus clustering technique.

By observing the results in the Table 6, one can establish the possible differences of the consensus clustering performances. Some technique features better performance than the others techniques, in conformity with their ARI and NMI values.

For some data sets, TEC.3 outperforms the others techniques whichever the base clustering algorithms, as D3-3g and D4-10gSS. For some others 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.

Based on the results of Table 6 and observing the comparison of the base clusterings variability established in the Table 5, we can affirm the following:

- ➢ Considering the simulated data set, D1-4g, for base clusterings obtained by SL the three techniques present differences. Actually, TEC.3 outperforms the others and we note that, SL presents statistically greater variability than the remaining hierarchical clustering.
- ➢ Regarding data set, D2-3g, whereas TEC.2 outperforms the others with base clusterings obtained by CL and TEC.3 outperforms the others, considering SL or AL. These clustering, statistically have the same variability as also greater than the remaining hierarchical clustering.
- ➢ For D2-3gr10, TEC.2 outperforms the others with base clusterings obtained by SL or AL, also TEC.3 outperforms the others, considering CL, which statistically have greater variability than the remaining clustering.
- ➢ As regard to D3-3gr10, TEC.3 outperforms the others techniques with base clusterings obtained by SL or CL or W clustering, which statistically have greater variability than AL clustering.
- ➢ Considering the real data set Iris, the TEC.2 outperforms the others techniques with base clusterings obtained by SL or AL clustering, besides, the TEC.3 features better performance than the other techniques, with CL and W, which, statistically have greater than the remaining clustering.
- ➢ Observing the data set Ecoli, TEC.1 has the best performance, relatively to the others, with AL and TEC.3 outperforms the others with CL or W which, have greater than the remaining clustering.
- ➢ For data set Wine, TEC.3 shows better performance than the others, with CL or W which have lower variability relatively the remaining clustering. While, the data set WDBC, TEC.3 shows better performance than the others with W which has also the lower variability relatively the remaining clustering.

Thus, in summary, TEC.3 of consensus clustering outperforms the others techniques, when it is applied to the hierarchical base clusterings having greater variability relatively to the others hierarchical base clusterings, notably for the data sets, D1-4g, D2-3g, D2-3gr10, D3-3gr10, Iris, and Ecoli. Also, TEC.2 prevails with hierarchical clustering having moderate variability, for the data sets D2-3g, D2-3gr10. For the data sets, D3-3g and D4-10gSS, TEC.3 outperforms the others techniques independently of the hierarchical base clusterings applied. About the data sets, Haberman's Survival, Breast Tissue and Blood, the three techniques show approximately the same performance for any of the hierarchical as base clusterings.

Thereby, we can assert that when there is differences on the performances of the consensus clustering techniques, TEC.3 has better performance, relatively to other techniques, independently of the hierarchical base clusterings used (it is observed for 2 data sets) or when it is applied to base clusterings with greater variability relatively to others (in these conditions there are 4 simulated data sets and 2 real data sets). The data sets excluded of the statements above have a known data clustering with overlapping clusters or have high dimensionality. Considering so, for some data sets tested, we may confirm the hypothesis under consideration, which the performance of some consensus clustering technique, as TEC.3, depends of the hierarchical base clusterings variance.

Table 6: Comparison of the consensus clustering's performances. The best relative results are highlighted.

			ARI			**NMI**		
	Data set	Clustering	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
	Haberman's 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*

V. CONCLUSIONS

In this paper we proposed to analyze empirically the clustering variability derived by the hierarchical algorithms, such as, Single Linkage, Complete Linkage, Average Linkage and Ward method, and from it, take knowledge about the performance of three techniques of consensus clustering, which are, Voting algorithm [8], Evidence Accumulation Clustering [9] and one based on Mutual Information and hyper graphs [13, 14]. Some data sets, synthetic and real, are used for this purpose. These performances were quantified considering measures by external criteria, applying the Adjust Rand index and the Normalized Mutual Information.

Through of these researches we search to define clustering's profiles achieved by the hierarchical algorithms according to their variability, and from that, decide which strategy of consensus clustering to apply.

These studies are performed by experimentally verify two hypotheses under consideration, one about, the difference of variability of the hierarchical clustering, wherein the analysis of their known properties led to the identification of a new property of these algorithms based on their variability. Another hypothesis studied, is the possibility of choosing the most appropriate consensus strategy, according to a particular type of clustering variances. Actually, when the consensus clustering techniques present different performances, in most of the cases the consensus technique based on Mutual Information and hyper graphs outperforms the others, with hierarchical clustering algorithm which have relatively higher variances.

REFERENCES

[1] K. Faceli and T.C. Sakata, *Multiple solutions in cluster analysis: partition x clusters*, Technical Report, 2016.

[2] R. Linden, *Técnicas de Agrupamento*, Revista de Sistemas de Informação da FSMA. N. 4, 2009, pp. 18-36. <http://www.fsma.edu.br/si/edicao4/FSMA_SI_2009_2_Tutorial.pdf>.

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

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

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

[6] A. Fred, *From Single Clustering to Ensemble Methods*, 2009.
<http://www.lx.it.pt/~afred/tutorials/D_Ensemble_Methods.pdf>.

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

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

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

[10] A. Fred and A. Lourenço, *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, 2008.

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

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

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

[14] S. T. Hadjitodorov, L. I. Kuncheva and L. P. Todorova, *Moderate diversity for better cluster*

ensembles, Information Fusion, vol. 7(3), 2006, pp. 264-275.

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

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

[17] I-Cheng Yeh, Department of Information Management , Chung-Hua University, Hsin Chu, Taiwan 30067, R.O.C, October 3, 2008. <https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center>.

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

[19] UCI Irvine Machine Learning Repository. <http://archive.ics.uci.edu/ml/datasets.html>.

[20] G. Karypis and V. Kumar, *A fast and high quality multilevel scheme for partitioning irregular graphs*, SIAM Journal of Scientific Computing, 20 (1), 2004, pp. 359-392.

[21] G. Karypis, R. Aggarwal, V. Kumar and S. Shekhar, *Multilevel hypergraph patitioning: Applications in VLSI domain*, in Proc. of the Design and Automation Conference, 1997.

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

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

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

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

[26] L. Sousa and J. Gama, *The application of hierarchical clustering algorithms for recognition using biometrics of the hand*, IJAERS, vol -1, Issue-7, Dec. 2014.

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

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

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

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

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

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

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

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

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

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