ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΤΙΤΛΟΣ

Clustering Validation Techniques on microarray data

ΧΡΥΣΑΝΘΗ ΙΩΑΝΝΗ ΠΑΠΑΓΙΑΝΝΑΚΟΠΟΥΛΟΥ

ΕΡΓΑΣΙΑ

Που υποβλήθηκε στο Τμήμα Στατιστικής

του Οικονομικού Πανεπιστημίου Αθηνών

ως μέρος των απαιτήσεων για την απόκτηση

Μεταπτυχιακού Διπλώματος

Συμπληρωματικής Ειδίκευσης στη Στατιστική

Μερικής Παρακολούθησης (Part-time)

Αθήνα

ΙΟΥΛΙΟΣ 2011

Abstract

The term cluster analysis encompasses a number of different algorithms and methods for grouping objects of similar kind into respective categories. A general question facing researchers in many areas of inquiry is how to organize observed data into meaningful structures, that is, to develop taxonomies. In other words cluster analysis is an exploratory data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. Given the above, cluster analysis can be used to discover structures in data without providing an explanation/interpretation. In other words, cluster analysis simply discovers structures in data without explaining why they exist.

While many clustering algorithms produce a candidate partitioning, relatively few attach a measure of confidence to the proposed clustering. An ideal clustering algorithm would do both; however, such a procedure is not always practical. As a result, the field of cluster validation attempts to remedy this by proposing methods to assess how well a proposed clustering of a dataset reflects its intrinsic structure.

The main goal of this thesis is to use validation indexes along with clustering algorithm using data from a microarray experiment. Once a clustering algorithm has been applied, validation indices were computed in order to conclude how many clusters the data set will have.

Περίληψη

Ο όρος ανάλυση κατά συστάδες εμπεριέχει έναν αριθμό διαφορετικών αλγορίθμων και μεθόδων για ομαδοποίηση αντικειμένων ίδιου είδους σε αντιπροσωπευτικές κατηγορίες. Ένα συχνό πρόβλημα που αντιμετωπίζουν οι ερευνητές είναι το πώς ορισμένα δεδομένα μπορούν να οργανωθούν σε δομές που έχουν νόημα. Με άλλα λόγια η ανάλυση κατά συστάδες είναι ένα ανιχνευτικό εργαλείο ανάλυσης δεδομένων, που βασικός του σκοπός είναι η κατηγοριοποίηση διαφορετικών αντικειμένων σε ομάδες, σε τέτοιο βαθμό ώστε ο βαθμός της σχέσης των αντικειμένων να είναι ο μέγιστος, αν ανήκουν στην ίδια ομάδα, και ο ελάχιστος σε άλλη περίπτωση. Παρόλα αυτά, η ανάλυση κατά συστάδες απλά ανακαλύπτει δομές στα δεδομένα, χωρίς να εξηγεί γιατί υπάρχουν αυτές οι δομές.

Παρότι πολλοί αλγόριθμοι ομαδοποίησης δεδομένων παράγουν ένα υποψήφιο διαχωρισμό δεδομένων, σχετικά λίγοι συμπεριλαμβάνουν ένα επίπεδο εμπιστοσύνης για το διαχωρισμό που έχει προταθεί. Ένας ιδανικός αλγόριθμος ομαδοποίησης θα έκανε και τα δύο. Παρόλα αυτά, μία τέτοια διαδικασία δεν είναι πάντα πρακτική. Σαν αποτέλεσμα, η τεκμηρίωση της ανάλυσης κατά συστάδες προσπαθεί να προτείνει μεθόδους για το πόσο καλά μία ομαδοποίηση ενός σετ δεδομένων αντανακλά την δομή του.

Ο κύριος σκοπός της εργασίας αυτής είναι η χρήση δεικτών τεκμηρίωσης μαζί με αλγορίθμους ομαδοποίησης σε δεδομένα από μικροσυστοιχίες, με σκοπό να καταλήξουμε στον αριθμό των ομάδων που θα μπορούσε να χωριστεί το σετ δεδομένων.

Table of Contents

Abstract I

Περίληψη II

Chapter 1 1

1.1 Introduction 1

1.2 Cluster Definition 1

1.2.1 Dissimilarity Measures 2

1.2.2 Linkage Methods 4

1.3 Hierarchical Clustering Algorithms 6

1.3.1 Agglomerative hierarchical nesting algorithm (AGNES) 6

1.3.2 Hierarchical Divisive Clustering (DIANA) 10

1.4 Partitioning Clustering Algorithms 11

1.4.1 K-Means Clustering 12

1.4.2. Partitioning Around Medoids 12

1.4.3 Clustering Large Applications (CLARA) 14

Chapter 2 17

2.1 Introduction 17

2.2. Monte Carlo use in cluster validity 19

2.3 External Criteria 22

2.3.1. Rand index 23

2.3.2. Jaccard coefficient 23

2.3.3. Fowlkes and Mallow’s index 24

2.3.4. Hubert’s Γ Statistic 24

2.3.5 Example of External Criteria 25

2.4 Internal Criteria 26

2.4.1 Davies-Bouldin Algorithm 26

2.4.2 Dunn Index 28

2.4.3 Silhouette Index 28

2.4.4 In Group Proportion Index (IGP) 30

2.4.5 Comparison of Internal Indices 31

2.5 Relative Criteria Measures 32

Chapter 3 33

3.1 Introductory Topics of Biology 33

    - 3.1.1 Genes 33
    - 3.1.2 Gene Expression 33
    - 3.1.3 Analysis of Gene Expression 34
    - 3.1.4 Microarrays – How do chips work 35
  - 3.2 Characteristics of Microarray Data 37
  - 3.3 Multivariate techniques in genetics. 39
  - 3.4 Experimental Results 39
    - 3.4.1. Data 40
    - 3.4.2 Clustering Algorithms Results 42
    - 3.4.3. Internal Criteria Results 50
- Chapter 4 57
  - 4. Conclusion 57
- Chapter 5 - Bibliography 59

Chapter 1

1.1 Introduction

Clustering may be defined as a process that aims to find partitions of similar objects. It is an unsupervised technique used to group together objects which are “close” to one another in a multidimensional feature space, usually for the purpose of uncovering some inherent structure which the data possesses **[1]**. An effective analysis should result in groups whose objects are homogeneous but at the same time objects of different groups should differ as much as possible **[2]**.

This chapter provides an overview about how the Clustering Algorithms work with examples and all the details needed to complete a clustering task. In Section 1.2, a mathematical definition for clustering is given. Dissimilarity measures are presented in section 1.2.1. Also, since we have to calculate the distance of one group with another group, we have to discuss about Linkage Methods. The latter are presented in Section 1.2.2.

There are two main types of clustering algorithms: 1) Hierarchical and 2) Partitioning discussed in Sections 1.3 and 1.4 respectively.

1.2 Cluster Definition

As a first approach, we need to define what a cluster is. Let X be our data set, that is,

$X = \{x_1, x_2, \dots, x_n\}$.

Now, let be the partition, R , of X into m sets, C_j, j=1,…,m. These sets are called clusters and need to satisfy the following two conditions **[3]:**

- $C_i \neq \emptyset , i = 1, \dots , m$
- $\bigcup_{i=1}^{m} C_i = X$

It is important to say that the objects (vectors) contained in a cluster C_i are more similar to each other and less similar to objects (vectors) contained in the other clusters. In order to join or separate vectors it is necessary to measure how similar, or dissimilar, two objects are. This task is carried out through the use of distances measures. Also we want to join, or separate, clusters, this can be done using Linkage Methods. Several distances measures and Linkage Methods are discussed in the next section.

1.2.1 Dissimilarity Measures

Dissimilarity measure the discrepancy between the two objects based on several features. Dissimilarity may also be viewed as measure of disorder between two objects. These features can be represented as coordinate of the object in the features space. There are many types of distance and similarity. Each similarity or dissimilarity has its own characteristics.

The dissimilarity coefficient, $d(i,j)$ are small when objects i and j are alike, otherwise, $d(i,j)$ become larger. The dissimilarity measures need to satisfy the following conditions:

- $0 \leq d(i,j) \leq 1$
- $d(i,i) = 0$
- $d(i,j) = d(j,i)$

Most of the clustering algorithms use dissimilarity measures to join, or to separate, objects. We will mention some of these measures:

- **City block Manhattan Distance**

  The taxicab metric is also known as city block distance, Manhattan distance, or Manhattan length, with corresponding variations in the name of the geometry. It examines the absolute differences between coordinates of a pair of objects. The taxicab distance, d_2, between two vectors p,q in an n-dimensional real vector space with fixed Cartesian coordinate system, is the sum of the lengths of the projections of the line segment between the points onto the coordinate axes. More formal,

$$d_2(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

Where x_i, and y_i are the i coordinates of x and y respectively, and x and y are objects of X. The City block distance is always greater than or equal to zero. The measurement would be zero for identical points and high for points that show little similarity.

- **Euclidean Distance**

The Euclidean Distance between points $\mathrm{x} = (\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{n})$ and $\mathrm{y} = (\mathrm{y}_1, \mathrm{y}_2, \dots, \mathrm{y}_\mathrm{n})$ is given by:

$$\mathrm{d}_1(\mathrm{x}, \mathrm{y}) = \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x}_\mathrm{i} - \mathrm{y}_\mathrm{i})^2}$$

where x_i, and y_i are the i coordinates of x and y respectively, and x and y are objects of X.

- **Mahalanobis Distance**

Formally, the Mahalanobis distance of a multivariate vector from a group of values with mean and covariance matrix S is defined as **[4]**:

$$\mathrm{d}(\mathrm{x}, \mathrm{y}) = \sqrt{(\mathrm{x} - \mathrm{y})^\mathrm{T} \mathrm{S}^{-1} (\mathrm{x} - \mathrm{y})}$$

If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance. Mahalanobis distance is widely used in cluster analysis and classification techniques.

- **Minkowski Distance**

The Minkowski distance is a metric on Euclidean space which can be considered as a generalization of both the Euclidean distance and the Manhattan distance. The

Minkowski distance between points $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$ is given by **[2][5]**:

$$d(x, y) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}$$

Minkowski distance is typically used with p being 1 or 2. The latter is the Euclidean distance, while the former is sometimes known as the Manhattan distance.

1.2.2 Linkage Methods

The Linkage methods are the quantitative measures used to join the two most similar clusters in the agglomerative clustering algorithm. In order to define the Linkage Methods, let $C_i$and C_j be two clusters, and let $|C_i|$ and$|C_j|$ denote the number of objects that each one have. Let $d(C_i, C_j)$ denote the dissimilarity measures between clusters C_i and C_j , and $d(i, j)$ the dissimilarity measure between two objects i, and j d where i is an object of C_i and j is an object of C_j . Some of the most used linkage methods are the following:

- **<u>UPGMA</u>**

  UPGMA (Unweighted Pair Group Method with Arithmetic Mean) is a simple agglomerative or hierarchical clustering method used in bioinformatics for the creation of phenetic phylogenetic trees (phenograms). The algorithm examines the structure present in a pairwise distance matrix (or a similarity matrix) to then construct a rooted tree (dendrogram). At each step, the nearest two clusters are combined into a higher-level cluster. The distance $d(C_i, C_j)$ between clusters C_i and C_j is defined as the average of all dissimilarities $d(i, j)$. That is **[6]**:

$$\delta(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{\substack{i \in C_i \\ j \in C_j}} d(i, j)$$

Single Linkage Clustering Method (SLINK)

SLINK is the short of single linkage clustering. Whereas the UPGMA clustering method defines the similarity between any two clusters as the arithmetic average of the similarities between the objects in one cluster and the objects in the other, SLINK does this differently. In SLINK, the distance between two clusters is taken to be the minimum of all the pairwise distances. Then, the SLINK is defined as follows:

$$\delta_2\left(C_i, C_j\right) = \min_{\substack{i \in C_i \\ j \in C_j}} d(i, j)$$

Complete Linkage Clustering Method (CLINK)

The CLINK is exactly the opposite of the SLINK. The CLINK is the maximum of all pairwise distances. It is defined as follows:

$$\delta_3\left(C_i, C_j\right) = \max_{\substack{i \in C_i \\ j \in C_j}} d(i, j)$$

Ward's Minimum Variance

In the Ward's Method the distance between two clusters is defined as a weighted version of the squared Euclidean distance of their mean vector. That is:

$$\delta_4^2\left(C_i, C_j\right) = \frac{|C_i|\left|C_j\right|}{|C_i| + \left|C_j\right|}\left|\left|\mu_\iota - \mu_j\right|\right|^2$$

where $\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$ and $\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$

Weighted pair-group Method using Arithmetic Averages (WPGMA)

The WPGMA is a variant of the UPGMA. The distance between clusters is calculated as a simple average. One starts with the original dissimilarities between objects and at each merger of clusters $\mathrm{C_i}$ and $\mathrm{C_j}$, forming some new cluster $\mathrm{C_k}$, the dissimilarities are updated by **[6]**:

$$\delta_5(C_k, C_m) = \frac{1}{2} d(C_i, C_m) + \frac{1}{2} d(C_j, C_m)$$

When there are unequal numbers of objects in the clusters, the distances in the original matrix do not contribute equally to the intermediate calculations, and the final result, is therefore, said to be weighted.

1.3 Hierarchical Clustering Algorithms

Hierarchical clustering groups data with a sequence of nested partitions, either from singleton clusters to a cluster including all individuals or vice versa. The former is known as agglomerative hierarchical clustering and the latter is called divisive hierarchical clustering. Both agglomerative and divisive clustering methods organize data into the hierarchical structure based on the proximity matrix. An important objective of hierarchical cluster analysis is to provide a picture of the data that can be easily interpreted, such as a dendrogram. A dendrogram lists the clustering one after another and cutting it at any level defines a clustering and identifies clusters.

Hierarchical clustering algorithms involve N −1 steps. It produces a hierarchy of nested clustering. At each step t, a new element is assigned to a cluster using the information produced in the previous step. Two categories of these algorithms are discussed: (1) Agglomerative (AGNES) and (2) Divisive (DIANA) hierarchical algorithms. Both algorithms have the disadvantage that once an element is assigned to a cluster there is no way to recover it later.

1.3.1 Agglomerative hierarchical nesting algorithm (AGNES)

The AGNES algorithm works by assigning each word to a separate cluster, and then iteratively joining together the most closely related (i.e. least dissimilar) clusters until a single super-cluster is formed. At first, each object is a small cluster by itself. Clusters are merged until only one large cluster remains which contains all the objects. At each stage the two "nearest" clusters are combined to form one larger cluster. The method is described below:

Let $d(C_i, C_j)$ be a function that measures the proximity between $C_i$andC_j, and t the current level of hierarchy. Then, the general scheme can be described as follows **[7]**:

a) At the beginning each of the objects in X forms a small cluster by itself.
b) At the first step, the two closest, or most similar, objects are joined using a dissimilarity measure $d(C_i, C_j)$ .That is, find the smallest value of the dissimilarity matrix and join the corresponding objects.
c) In the second step we have $N - 1$ clusters. Now we will want to merge the closest clusters using one of the linkage method previously described.
d) At step t we have $N - (t - 1)$ clusters, and we want to join the closest clusters as in the previous step.
e) Repeat until all the vectors lie in a single cluster.

The following example **[8]** shows how the AGNES algorithm works:

In Table 1.1, a data set consisting of flowers is given **[6]**. 8 variables have been measured on these flowers: winters, shadow, tubers, color, soil, preference, height, and planting distance. Before we begin with the steps of the algorithm it is necessary to standardize the data.

Flowers	win	shadow	tuber	color	soil	prefer	height	dist
Myosotis	0	1	0	5	2	2	20	15
Iris	1	1	1	5	3	8	45	10
Lily	1	1	1	1	2	9	90	25
Red Rose	1	0	0	4	2	18	200	60
Scotch Rose	1	0	0	2	2	17	150	60
Tulip	0	0	1	2	1	5	25	10

Table 1.1 – Standardized Data: Flowers

Step 1

Once the data has been standardized, the next step is to find the Proximity Matrix in order to find the nearest object among X. That is, find the Euclidean Distance among all the possible pairs of vectors (Table 1.2). Looking at the Dissimilarity Matrix

(Table 2.2) the nearest objects are Red Rose and Scotch Rose (d = 1.36). Then, we join it and now we have five clusters: (1) {Red Rose, Scotch Rose}, (2) Myosotis, (3) Iris, (4) Lily and (5) Tulip.

	Myosotis	Iris	Lily	Red Rose	Scotch Rose	Tulip
Myosotis	0					
Iris	3.26	0				
Lily	3.84	2.95	0			
Red Rose	4.82	4.56	4.01	0		
Scotch Rose	4.72	4.54	3.37	1.36	0	
Tulip	3.53	4.52	3.39	5.01	4.51	0

Table 1.2 - Distances

Step 2

In this step we needed to join the nearest clusters, to do this we will use a linkage method, that is, find the distance between the clusters. We will calculate the new distances using equation of CLINK and will obtain a new matrix. The new distances are:

$$d(\{red, scotch\}, myosotis) = \max\big(d(red, myosotis), d(scotch, myosotis)\big)$$

$$= max 4.82,4.71) = 4.82$$

$$d(\{red, scotch\}, iris) = \max(4.56,4.53) = 4.56$$

$$d(\{red, scotch\}, lily) = \max(4.01,3.37) = 4.01$$

$$d(\{red, scotch\}, tulip) = \max(5.01,4.51) = 5.01$$

The new proximity matrix is:

	{Red,Scotch}	Myosotis	Iris	Lily	Tulip
{Red,Scotch}	0	4.82	4.56	4.01	5.01
Myosotis	4.82	0	3.26	3.84	3.53
Iris	3.26	3.26	0	2.95	4.52
Lily	3.84	2.95	2.95	0	3.39
Tulip	4.82	4.56	4.52	3.39	0

Table 1.3 - Distances

The nearest clusters are Iris and Lily and therefore the new distances are:

$$d(\{iris, lily\}, \{red, scotch\}) = \max(4.56, 4.01) = 4.56$$

$$d(\{iris, lily\}, \{tulip\}) = \max(4.52,3.39) = 4.52$$

$$d(\{iris, lily\}, \{myosotis\}) = \max(3.84,3.26) = 3.84$$

Now the new proximity matrix is:

	{Iris,Lily}	**{red,scotch}**	**Myosotis**	**Tulip**
{Iris,Lily}	0	4.56	3.84	4.52
{red,scotch}	4.56	0	4.82	5.01
Myosotis	3.84	4.82	0	3.53
Tulip	4.52	5.01	3.53	0

Table 1.4 - Distances

As in the previous step we need to join the nearest clusters, that is, myosotis and tulip. Then the new distances will be:

$$d(\{iris, lily\}, \{myosotis, tulip\}) = \max(3.84, 4.52) = 4.52$$

$$d(\{myosotis, tulip\}, \{red, scotch\}) = \max(4.82, 5.01) = 5.01$$

The last proximity matrix we have is:

	{Iris,Lily}	**{red,scotch}**	**{myosotis,tulip}**
{Iris,Lily}	0	4.56	5.01
{red,scotch}	4.56	0	4.52
{Myosotis, tulip}	5.01	4.52	0

Table 1.5 - Distances

The new clusters are now: (1) {red, scotch, myosotis, tulip} and (2) {iris, lily}.

Step 4

This is the last step and the only thing to do is to join the two clusters. Then, in the last step all the vectors lie in a single cluster. Due to the fact that AGNES algorithm is an hierarchical one, the only way to obtain a graph of results is a dendrogram. For a dendrogram of the example above see Diagram 1.1.

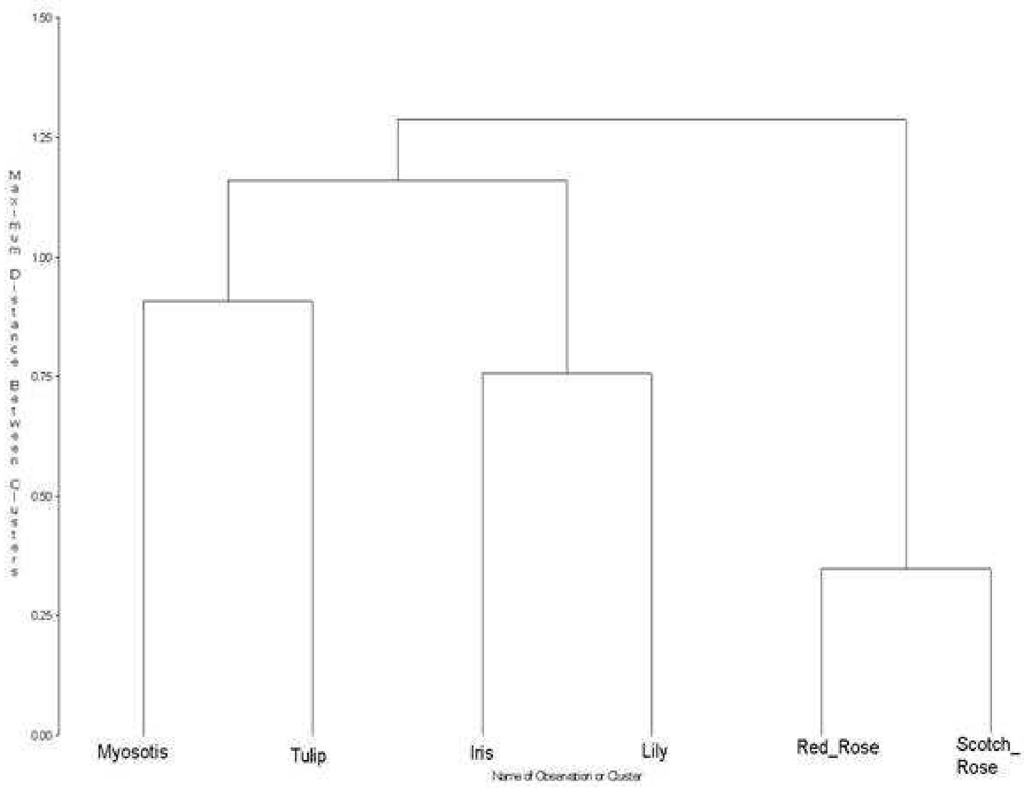


Diagram 1.1 – Clustering Tree of AGNES algorithm

1.3.2 Hierarchical Divisive Clustering (DIANA)

Hierarchical Divisive Algorithms starts with a single cluster of all the given objects and keep splitting the clusters based on a dissimilarity measure to obtain a partition of singleton clusters **[6]**. The algorithm can be described as follows:

(a) Before starting the algorithm all objects in X are together in a single cluster.

(b) At the first step, split the data set into two clusters. For this purpose, look for the object for which the average dissimilarity to all other objects is largest. The object with the largest dissimilarity initiate a new cluster, named the splinter group.

(c) For each object in the larger group, compute the average dissimilarity with the remaining objects, and compare it to the average dissimilarity with the objects of the splinter group. The object in the larger group with the largest difference changes sides; it is moved to the splinter group. Repeat the computations until all the differences are negatives.

(d) At the next step, divide the biggest cluster, that is, the cluster with the largest diameter. The procedure is the same as in the previous step.

(e) In the following steps, divide the biggest cluster following the previous step.

(f) The process continues until all objects form a singleton.

1.4 Partitioning Clustering Algorithms

A partitioned clustering obtains a single partition of the dataset instead of a clustering structure, such as the dendrogram. The algorithm used in the Partitioning Clustering is based on the search of k representative objects among the objects of the data set. These k-representative objects should represent various aspects of the structure of the data, and are often called centrotypes. After finding a set of k representative objects, the k clusters are constructed by assigning each object of the dataset to the nearest representative object. Below we point out some of the advantages and disadvantages of partitioning clustering algorithms:

Advantages

- With a large number of variables, partitioning clustering algorithms may be computationally faster than hierarchical clustering (if k is small).
- Partitioning clustering algorithms may produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

Disadvantages

- Difficulty in comparing quality of the clusters produced (e.g. for different initial partitions or values of K affect outcome).
- Fixed number of clusters can make it difficult to predict what K should be.
- Different initial partitions can result in different final clusters. It is helpful to rerun the program using the same as well as different K values, to compare the results achieved.

In this section we discuss two portioning clustering algorithms: K-means and Partitioning Around Medoids.

1.4.1 K-Means Clustering

The K-means algorithm is a popular non-hierarchical technique. The algorithm proceeds as follows **[10]**:

1. Select K – points as initial centroids.
2. Assign each observation to its closest cluster centroid. That generates a new partition.
3. Compute the centroid of the new partition.
4. Recompute the centroid of each cluster until centroids do not change.

Some important characteristics of the K-means algorithm are: (1) it is computationally fast, (2) it is sensible to outliers since an object with an extremely large value may substantially distort the distribution of data, and (3) can be performed with missing values. PAM that follows is a generalization of the K-means clustering algorithm.

1.4.2. Partitioning Around Medoids

PAM (Partitioning Around Medoids) was developed by Kaufman and Rousseeuw **[6]**. To find k clusters PAM's approach is to determine a representative object for each cluster. This representative object, called a medoid, is meant to be the most centrally located object within the cluster. Once the medoids have been selected each non-selected object is grouped with the medoid to which it is the most similar.

The algorithm is divided into two phases. In the first phase, called BUILT, the k representative objects are chosen. The second phase, called SWAP, is attempted to improve the quality of the clustering. The algorithm is the following [6][7]:

BUILT PHASE

In this phase the first object chosen is the one for which the sum of the dissimilarities to the other objects is the smallest. This object is the most centrally located in the set of objects. At each step the object that decreases the objective function is selected.

1. Consider an object i which has not yet been selected.
2. Consider a non selected object j and calculate it's dissimilarity D_j with the first object chosen and calculate it's dissimilarity $d(i,j)$ with object i. Calculate the difference between D_j and. If the difference is positive, then object j will contribute in the selection of object i. Then calculate,

$$C_{ij} = max\left(D_j - d(i,j), 0\right)$$

3. Calculate the total gain obtained if object i is selected,

$$\sum_j C_{ij}$$

4. Choose the object that maximizes

$$\sum_j C_{ij}$$

The process ends when the k representative objects have been found. Now, consider all the pair on objects (i, h) for which object i have been selected, but object h has not. The main objective is to determine if there is a positive effect when a swap is carried out, that is, when object i is no longer selected as a representative object, but object h is.

SWAP PHASE

To calculate the effect of a swap between objects i and h the following calculations need to be completed.

1. First, consider an object j that has not been selected. Then calculate its contribution C_{jih} to the swap:

   a. If j is near from one of the other representative objects than from both i and h then the contribution of object j to the swap is C_{jih} = 0.

   b. Consider this two situations if j is not further from i than from any other selected representative object$\big(d(j,i) = D_j\big)$.

   b1. If j is closer to h that from any other representative object, that is, $d(j.h) < E_j$ where E_j is the dissimilarity between j and the second most similar representative object, then the contribution of object j to the swap is

$$C_{jih} = d(j,h) - d(j,i)$$

   b2. If j is at least as distant from h as from the second closest representative object, that is, $d(j.h) \geq E_j$ then the contribution of object j to the swap is $C_{jih} = E_j - d(j.i)$

   c. If j is more distant from object i that from at least one of the other representative object the contribution to the swap is:

$$C_{jih} = d(j,h) - d(j.i)$$

2. Secondly, add the contributions $\boldsymbol{C_{jih}}$ to calculate the total result of the swap:

$$\boldsymbol{T_{ih} = \sum_{j} C_{jih}}$$

3. The next step will be to select the pair $(\boldsymbol{i.h})$ which minimizes.

The swap is carried out if minimum is negative and the algorithm return to step1. If minimum is positive or zero then the swap is not carried out.

1.4.3 Clustering Large Applications (CLARA)

Instead of finding representative objects for the entire data set, CLARA draws a sample of the data set, applies PAM on the sample, and finds the medoids of the sample. The point is that if the sample is drawn in a sufficiently random way, the medoids of the sample would approximate the medoids of the entire data set. To come up with better approximations, CLARA draws multiple samples and gives the best clustering as the output. The algorithm of CLARA is presented below **[10]**:

1. For $\boldsymbol{i = 1}$ to 5, repeat the following steps:
2. Draw a sample of 40 + 2k objects randomly from the entire data set l, and call Algorithm PAM to find k medoids of the sample.
3. For each object O_j in the entire data set, determine which of the k medoids the most similar is to O_j.
4. Calculate the average dissimilarity of the clustering obtained in the previous step. If this value is less than the current minimum, use this value as the current minimum, and retain the k medoids found in Step (2) as the best set of medoids obtained so far.
5. Return to Step (1) to start the next iteration.

Complementary to PAM, CLARA performs satisfactorily for large datasets.

Chapter 2

2.1 Introduction

Clustering algorithms generally rely on some prior knowledge of the structure present in a data set. Clustering applied to a data set with no naturally occurring clusters will impose artificial and meaningless structure. The procedure that consists in examining a data set to determine if structure is actually present and thus determine if clustering is worthwhile operation is a poorly investigated problem known as clustering tendency.

Once we assume that X possesses a clustering structure we want to unreveal it. Since the clustering results are not completely reliable, it is necessary further evaluation of these resulting clustering. Cluster Validity is the procedure of evaluating, quantitatively, the results of a clustering algorithm **[11]**.

The aim of the cluster validity is to find the partitioning that best fits the underlying data. Usually 2D data sets are used for evaluating clustering algorithms as the reader easily can verify the result. But in case of high dimensional data the visualization and visual validation is not a trivial task therefore some formal methods are needed **[12]**.

The process of evaluating the results of a clustering algorithm is called cluster validity assessment. Two measurement criteria have been proposed for evaluating and selecting an optimal clustering scheme **[12]**:

- **Compactness**: The member of each cluster should be as close to each other as possible. A common measure of compactness is the variance.

- **Separation**: The clusters themselves should be widely separated. There are three common approaches measuring the distance between two different clusters: distance between the closest member of the clusters, distance between the most distant members and distance between the centers of the clusters.

In cases of biological data, the use of prior biological knowledge and assumptions may be necessary and important in the final interpretation of a cluster analysis. However, this process of data analysis is highly subjective, and may be a dangerous endeavor. In particular, researchers may unwittingly overrate clusters that reinforce their own assumptions, and ignore surprising or contradictory results. Therefore, it is not an acceptable means of replacing an unsupervised validation step, in which the significance of individual clusters in terms of the underlying data distribution is verified.

The fact that a validation step is needed follows from the following two issues that arise when using clustering algorithms **[13]**:

- **Bias of clustering algorithms towards particular cluster properties**. Clustering algorithms are biased towards partitions that are in accordance with their own clustering criterion. This is at the bottom of the fundamental discrepancies observable between the solutions produced by different algorithms.

- **Non-significance of results in the absence of natural clusters**. Unsupervised classification relies on the existence of a distinct structure within the data. However, most clustering algorithms return a clustering even in the absence of actual structure, leaving it to the user to detect the lack of significance of the results returned.

In General terms there are three approaches to investigate cluster validity **[14]**:

- **External Criteria**: this implies that we evaluate the results of a clustering algorithm based on a pre-specified structure, which is imposed on a data set and reflects our intuition about the clustering structure of the dataset. It is applicable when external information like class labels are available.

- **Internal Criteria**: we may evaluate results of a clustering algorithm in terms of quantities that involve the vectors of the dataset themselves. An internal criteria is an independently meaningful measure of the cluster/validity, that can be computed given nothing but the data and the clustering.

- **Relative Criteria**: the basic idea is the evaluation of a clustering structure by comparing it to other clustering schemes, resulting by the same algorithm but with different parameter values.

Both internal and external criteria are based on statistical methods and they have high computation demand. The external validity methods evaluate the clustering based on some user specific intuition. The internal criteria are based on some metrics which are based on data set and the clustering schema. The main disadvantage of these two methods is its computational complexity. The basis of the relative criteria is the comparison of the different clustering schema. One or more clustering algorithms are executed multiple times with different input parameters on same data set. The aim of the relative criteria is to choose the best clustering schema from the different results. The basis of the comparison is the validity index. Several validity indices have been developed and introduced.

2.2. Monte Carlo use in cluster validity

When Hypothesis Testing is done in Cluster Validation, the null Hypothesis H_0 consists in testing whether the data of X possess a random structure or not. Thus, the null hypothesis should be a statement of randomness concerning the structure of X.

$$H_0: Data\ X\ possess\ \ a\ random\ structure$$
$$H_A: Data\ X\ does\ not\ possess\ a\ random\ structure$$

Also let $\overline{D}_\rho$ be the critical interval corresponding to significance level ρ of a test statisticq_i and Θ the set of all possible values that θ may take under hypothesis H_A.

The power function of the test is defined as [7]:

$$W(\theta) = P\left(q \in \bar{D}_p \middle| \theta \in \Theta\right)$$

For a specific $\theta \in \Theta$, $W(\theta)$ is known as the test power under the alternative θ. In words is the probability that q lies in the critical region when the value of the parameter vector is θ. This is the probability of making the correct decision when H_0 is rejected. The power function can be used for the comparison of two different statistical tests. The test whose power under the alternative hypotheses is greater is always preferred.

There are two types of errors associated with a statistical test.

- **Type I Error**: Suppose that H_0 is true. If $q_x \in \bar{D}_p$, H_0 will be rejected even if it is true. The probability of such error is ρ. The probability of accepting H_0 when it is true is $1 - \rho$.
- **Type II Error**: Suppose that H_0 is false. If $q_x \in \bar{D}_p$, H_0 will be accepted even if it is false. The probability of such error is $1 - W(\theta)$ and it depends on the specific value of θ.

The goal of using Monte Carlo techniques is the computation of the probability density function. First a large amount of datasets is generated by a normal distribution. For each one of the synthetic datasets called $\mathrm{X_i}$, the value of the defined index denoted as $\mathrm{q_i}$ is computed. Then based on the respective values of $\mathrm{q_i}$ for each of the datasets $\mathrm{X_i}$, we create a scatter plot. This scatter plot is an approximation of the probability density function of the index.

We present here a Monte Carlo algorithm [15]:

1. **For** $i = 1\ to\ r$ **do**:
2. Generate randomly from a distribution a data $\mathrm{X_i}$ with N vectors (points) in the area of X
3. Assign each vector $y_{j,i}$ of $\mathrm{X_i}$ to the group that $\mathrm{x_j} \in \mathrm{X}$ belongs, according to the partition P.
4. Run the same clustering algorithm used to produce C, for each $\mathrm{X_i}$ and let C_i the resulting clustering structure.

```
5. Compute $q(C_i)$ value of the defined index $q$ for P and $C_i$.
6. End for
7. Create scatter plot of the r validity index values,.
```

There are three different possible cases depending on the critical interval, corresponding to significant level ρ. The probability density function of a statistic index q, under H_0 has a single maximum and the $\bar{D}_p$ region is either half line or a union of two half lines. Assuming that the scatter plot has been generated using r values of the index$q$, called $\mathrm{q_i}$, in order to accept or reject the null hypothesis we examine the following conditions:

- If the shape is right tailed then Reject H_0 else Accept
- If the shape is left tailed then Reject H_0 else Accept
- If the shape is two tailed then Accept H_0.

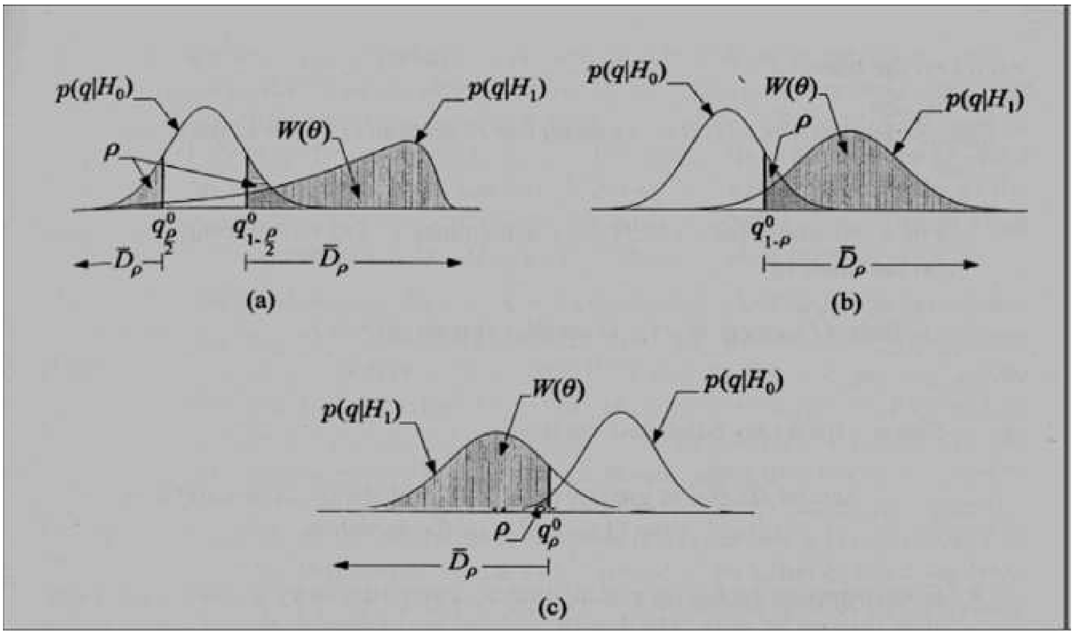


Figure 2.1 – Probability Density Function [6]

2.3 External Criteria

The clustering validation using external criteria is based on the null hypothesis, which represents a random structure of a dataset **[16]**. It evaluates the resulting clustering structure by comparing it to an independent partition of the data built according to the null hypothesis of the dataset. This kind of test leads to high computation costs. Generally the Monte Carlo techniques are suitable for the high computation problem and generate the needed probability density function.

Let us define a clustering structure C and a defined partition, P, before we can apply the cluster validation technique. We consider a clustering,C that result from a specific clustering algorithm, and compare it with a independently drawn partition P of X. Suppose that $C = \{C_1, \dots, C_m\}$ and. The number of clusters in C and the partition in P do not need to be the same.

Consider the following pair of vectors(x_u, x_v). Then we refer to it depending whether or not this pair of vectors belongs to the same cluster or partition.

Let us define the following notation **[12]**:

- SS if both vectors belong to the same cluster in C and to the same group in$P$.

- SD if both vectors belong to the same cluster in C and to different groups in$P$.
- DS if both vectors belong to different clusters in C and to the same group in$P$.
- DD if both vectors belong to different clusters in C and to different groups in$P$.

Then let's define that a, b, c and d are the numbers of SS, SD, DS and DD respectively, then $a + b + c + d = M$ which is the maximum number of all pairs in the dataset. Using the above we can define the following external indices to measure the degree of similarity between C and$P$.

2.3.1. Rand index

The rand index R measures how closely the clusters created by the clustering algorithm match the ground truth. It produces measures with values in the interval [0,1] with 1 meaning a perfect match between the result of clustering algorithm and the real clustering pattern **[17]**. It is defined as **[12]**:

$$\mathrm{R} = \frac{\mathrm{a} + \mathrm{d}}{\mathrm{M}}$$

where $(a + d)$ is the sum of SS pairs of vector plus the DD pairs. The values of this index lie between 0 and 1, and values close to 1 indicates high agreement between C and$P$.

However there are some known problems with Rand Index such as the fact that the expected value of the Rand Index of two random partitions does not take a constant value (say zero) or that the Rand index approaches its upper limit of unity as the number of clusters increases. In order to overcome these limitations ***Adjusted Rand Index*** has been created **[18]**.

In fact Adjusted Rand index became one of the most successful cluster validation indices and it is recommended as the index of choice for measuring agreement between two partitions in clustering analysis with different numbers of clusters. Adjusted Rand Index can be computed as:

$$ARI = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c) + (c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c) + (c+d)(b+d)]}$$

2.3.2. Jaccard coefficient

The Jaccard Coefficient measures the proportion of pairs that are in the same cluster and in the same partition from those that are either in the same cluster or in the same partition. In other words, it is the ratio of the number of positive matches to the total number of characters minus the number of negative matches. It is defined as follows [12]:

$$J = \frac{a}{a+b+c}$$

where $a + b + c \ = \ SS + SD + DS$. As in the Rand Index, the values of this coefficient lie between 0 and 1, and values close to 1 indicate high agreement between C and P.

2.3.3. Fowlkes and Mallow's index

The Fowlkes-Mallows Index is the geometrical mean of two probabilities: the probability that two random objects are in the same cluster given they are in the same group, and the probability that two random objects are in the same group given they in the same cluster [19]. The FM index is defined as below:

$$FM = {}^{a}\!/\!\sqrt{m_1 m_2} = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$$

As in the Rand Index and Jaccard Coefficient, values close to 1 indicate high agreement between C and P.

2.3.4. Hubert's Γ Statistic

The Hubert's Γ Statistic measures the correlation between the matrices, X and Y , of dimension $N \ \times \ N$, drawn independently of each other, where $X(i,j)$ equals to 1 if the pair of vectors (x_i, x_j) belong to the same cluster in C and 0 otherwise, and $Y\ (i,j)$ equals to 1 if the pair of vector (x_i, x_j) belongs to the same group in P and 0 otherwise. The statistic is defined as follows [7]:

$$\Gamma = (1/M) \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} X(i,j)Y(i,j)$$

High values of this index indicate a strong similarity between X and Y.

It might be more useful to have values of -1 to 1 range and therefore ***Normalized Γ statistic*** is used **[7]**:

$$\hat{\Gamma} = \frac{\left[(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(X(i,j)-\mu_x)\left(Y(i,j)-\mu_y\right)\right]}{\sigma_x\sigma_y}$$

where $X(i,j)$ and $Y(i,j)$are the (i,j) elements of the matrices X,Y respectively that we have to compare. Also μ_x,μ_y,σ_x,σ_y are the respective means and variances of X,Y matrices. The last index takes values between -1 and 1.

2.3.5 Example of External Criteria

In order to show how these indices are calculated let us make an example. The general form of the example we will show is a contingency table:

Class\Cluster	$\boldsymbol{v_1}$	$\boldsymbol{v_2}$	...	$\boldsymbol{v_C}$	***Sums***
u_1	n_{11}	n_{12}	...	n_{1C}	$n_{1.}$
.	.	.	...	.	.
.	.	.		.	.
.	.	.		.	.
u_R	n_{R1}	n_{R2}	...	n_{RC}	$n_{R.}$
Sums	$\boldsymbol{n_{.1}}$	$\boldsymbol{n_{.2}}$	...	$\boldsymbol{n_{.C}}$	$\boldsymbol{n_{..}}$

Table 2.1 –Contingency Table

where n_{ij} the number of objects that are in both cluster $\boldsymbol{v_1}$ and class u_1.

Table 2.2 is a contingency table in the same form as Table 2.1 **[20]**:

Class\Cluster	$\boldsymbol{v_1}$	$\boldsymbol{v_2}$	$\boldsymbol{v_3}$	***Sums***
u_1	1	1	0	**2**
v_2	1	2	1	**4**
v_3	0	0	4	**4**
Sums	**2**	**3**	**5**	**10**

Table 2.2 – Contingency Table

According to table we have:

$a = 7$

$$b = 6$$

$$c = 7$$

$$d = 25$$

Therefore we have:

- Rand Index = 0.711
- Adjusted Rand Index = 0.313
- Jaccard = 0.35
- Fowlkes and Mallows = 0.519
- Hubert = 0.313

2.4 Internal Criteria

Contrary to external criteria, internal validation is based on the information intrinsic to the data alone. We may evaluate the results of a clustering algorithm using information that involves the vectors of the datasets themselves. Internal criteria can roughly be subdivided into two groups: the one that assesses the fit between the data and the expected structure and others that focus on the stability of the solution **[21]**. In the following section, we present an overview of internal validity indexes:

2.4.1 Davies-Bouldin Algorithm

Let s_i be measure of dispersion of cluster C_i and $d(C_i, C_j) \equiv d_{ij}$ the dissimilarity between two clusters. A similarity index R_{ij} between C_i and C_j satisfy the following **[3]**:

- $R_{ij} \geq 0$
- $R_{ij} = R_{ji}$
- If $s_i = 0$ and $s_j = 0$ then $R_{ij} = 0$
- If $s_j > s_k$ and $d_{ij} = d_{ik}$ then $R_{ij} > R_{ik}$
- If $s_j = s_k$ and $d_{ij} < d_{ik}$ then $R_{ij} > R_{ik}$

These conditions state that R_{ij} is nonnegative and symmetric. A choice for a that satisfies these conditions is **[6]**:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

Then the Davies-Bouldin index is defined as:

$$DB_m = \frac{1}{m}\sum_{i=1}^{m} R_i$$

where $R_i = max_{j=1,\ldots,m j \neq i} R_{ij}$, $i = 1, \ldots, m$

The dissimilarity between clusters C_i and C_j, in a l-dimensional space is defined as:

$$d_{ij} = \left\|\bar{x}_i - \bar{x}_j\right\| = \sqrt{\sum_{k=1}^{l} \left|\bar{x}_{ik} - \bar{x}_{jk}\right|^2}$$

And the dispersion of a cluster C_i is defined as:

$$s_i = \sqrt{\frac{1}{n_i}\sum_{x \in C_i} |x - \bar{x}_i|^2}$$

The DB_m is the average similarity between each cluster and its most similar one. Small values of DB correspond to clusters that are compact, and whose centers are far away from each other. Consequently, the number of clusters that minimizes DB is taken as the optimal number of clusters.

2.4.2 Dunn Index

The Dunn index is defined as **[22]**:

$$D_m = \min_{i=1,\dots,m} \left\{ \min_{j=i+1,\dots,m} \left(\frac{d(C_i, C_j)}{\max_{k=1,..,m} diam(C_k)} \right) \right\}$$

where the dissimilarity function between two clusters C_i and C_j is:

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y)$$

and the diameter of a cluster C is defined as :

$$diam(C) = \max_{x,y \in C} d(x, y)$$

If X contains compact and well-separated clusters Dunn Index will be large and diameter of the cluster is expected to be small.

2.4.3 Silhouette Index

The silhouette index is useful when it is seeking compact and clearly separated clusters. In order to construct silhouettes we need a partition obtained by the application of some clustering algorithms, and the proximity matrix containing all the proximities between objects.

For a given cluster, this method assigns to each object of the cluster a quantitative measure, known as the silhouette width **[23]**. The silhouette width indicates the membership of object i in the cluster it has been assigned. Let i any object in the data set, and denote by C_j the cluster to which object i has been assigned. Let $a(i)$ the average dissimilarity between i and all the other objects in clusterC_j. Consider any cluster C_k different to cluster C_j, and compute $b(i) = \min d(i, C_k)\, , k = 1,2, \dots, c\ ; k \neq j$. Then the silhouette width is:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

A neighbor of object i is the cluster C_k for which the minimum is obtained, that is, $d(i, C_k) = b(i)$. Cluster C_k represents the second best choice for object i.

From the definition we can see that$-1 < s(i) < 1$. A value of $s(i)$ close to 1 is obtained when the within dissimilarity $a(i)$ is much smaller than the smallest between dissimilarity $b(i)$. Therefore we can say that object i is well clustered. On the other hand, if $s(i)$ take values close to -1 implies that $a(i)$ is much larger that $b(i)$. In this case we can say that object i has been misclassified, so object i may be reassigned. If $a(i)$ and $b(i)$ have similar values then $s(i)$ is about zero. In this situation object i lies equally far away from both cluster C_j and C_k.

Having computed the silhouette width for each object we can construct a graphical display **[24]**. The silhouette shows which objects lie well within their cluster, and which one are merely somewhere in between clusters. A wide silhouette indicates large $s(i)$ values, and hence a pronounced cluster. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters. This measure ranges from +1, indicating points that are very distant from neighboring clusters, through 0, indicating points that are not distinctly in one cluster or another, to -1, indicating points that are probably assigned to the wrong cluster. Below there is an example of silhouettes width plot (Figure 2.2) **[25]**:

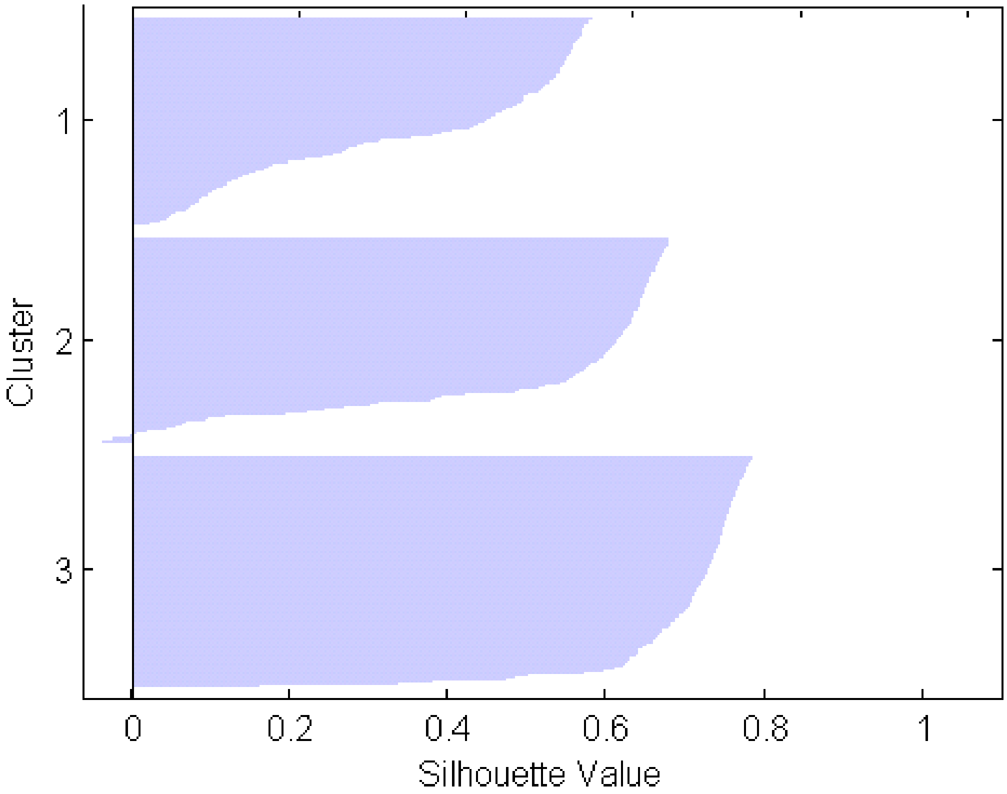


Figure 2.2 – Silhouette Width Plot

From the silhouette plot, we observe that most points in the third cluster have a large silhouette value, greater than 0.6, indicating that the cluster is somewhat separated from neighboring clusters. However, the first cluster contains many points with low silhouette values, and the second contains a few points with negative values, indicating that those two clusters are not well separated.

2.4.4 In Group Proportion Index (IGP)

The In-Group Proportion is the proportion of observations in a cluster whose nearest neighbors are in the same cluster. IGP captures the idea of prediction accuracy and quantifies the degree to which points close to each other are predicted to belong to the same cluster. It is computed as **[26]**:

$$IGP(C_k) = \frac{n(i: i \in C_k \, and \; i^* \in C_k)}{n(C_k)}$$

where $i^* = argmin\, d(i, x)$ in which d is a distance function. IGP scores take values between 0 and 1 with larger scores indicating a better predictive ability.

2.4.5 Comparison of Internal Indices

After mentioning the most important of the internal indices for clustering validation, it is necessary to identify which one is more accurate and at the same time make a general comparison.

The results from several experiments **[27]** have shown that the Silhouette index produces more accurate results than the Davies-Bouldin index. However, the time complexity of the Silhouette index computation is much greater than the time complexity of the Davies-Bouldin index computation. Thus the Davies-Bouldin index has a great advantage over the Silhouette index, regarding the overall performance.

At the same time another experiment **[28]** shows that best results were obtained using the Silhouette Width followed by the Dunn-index and Davies-Bouldin index. Given the noisy nature of biological data, robust measures like the Silhouette Width are preferable to noise-sensitive measures like the Dunn index, which is instable against outliers due to the consideration of only two distances. The Davies-Bouldin index requires the computation of the cluster centre, which cannot be achieved by average determination when dealing with binary data. An inappropriate choice of method for cluster center determination might have been one of the reasons for the insufficient clustering results obtained by this distance measure.

If it is not so clear which index is the appropriate, then combination of these methods may be successfully used for the assessment of cluster validity **[29]**. Normalization and weighed voting techniques are proposed to improve the prediction of the number of clusters based on multiple indices. Normalization allows smoothing the effect of the highest values on the calculation of the average index values. Moreover, it effectively highlights the differences between the average index values from different clustering configurations.

Finally, Kapp and Tibshirani in **[30]** propose that of the cluster quality measures considered, the IGP was the best at quantifying how likely a point was to be assigned to a different cluster.

2.5 Relative Criteria Measures

The relative criteria does not involve statistical test as in the two criteria discussed above. In this case the main idea is to choose, from a set of clustering, the best one according to a pre-specified criterion. Let A be the set of parameter associated with a specific algorithm. For example, some algorithm has the number of cluster nc as a parameter. The problem can be stated as: Among the clustering obtained by a specific clustering algorithm, for different values of the parameter, choose the one that best fits the data set$X$. Consider the following cases **[3]**:

- *A does not contain the number of clusters,*nc*, as a parameter.*

  The choice of the appropriate parameter values for this type of algorithm is based on the assumption that if X possesses a clustering structure, then a large range of values of the parameters in A can capture such a structure. Then, run the algorithm for a wide range of values for nc, and choose the largest range for which nc remains constant. The appropriate value for nc is the values that correspond to the middle to the range.

- *A contains the number of clusters,* nc *as a parameter.*

  First select a suitable index q. Run the clustering algorithm for all values on nc between $ncmax$ and $ncmin$, chosen a priori. For each value of nc, run the algorithm n times, using different set of values for the parameters in A. Plot the best values of q, obtained for each nc, versus nc. The values of q in where a maximum and a minimum are obtained indicate good clustering.

Chapter 3

3.1 Introductory Topics of Biology

3.1.1 Genes

A gene is a unit of heredity in a living organism. Living things depend on genes, as they specify all proteins and functional RNA chains. Genes hold the information to build and maintain an organism's cells and pass genetic traits to offspring. All organisms have many genes corresponding to many different biological traits, some of which are immediately visible, such as eye color or number of limbs, and some of which are not, such as blood type or increased risk for specific diseases, or the thousands of basic biochemical processes that comprise life. The vast majority of living organisms encode their genes in long strands of DNA.

A modern working definition of a gene is "*a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and or other functional sequence regions* ". Colloquial usage of the term *gene* (e.g. "good genes", "hair color gene") may actually refer to an allele: a *gene* is the basic instruction, a sequence of nucleic acids (DNA or, in the case of certain viruses RNA), while an *allele* is one variant of that gene. Thus, when the mainstream press refers to "having" a "gene" for a specific trait, this is generally inaccurate. In most cases, all people would have a gene for the trait in question, but certain people will have a specific allele of that gene, which results in the trait variant. In the simplest case, the phenotypic variation observed may be caused by a single letter of the genetic code - a single nucleotide polymorphism **[31]**.

3.1.2 Gene Expression

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. These products are often proteins, but in non-protein coding genes such as rRNA genes or tRNA genes, the product is a functional RNA. The process of gene expression is used by all known life - eukaryotes (including multicellular organisms), prokaryotes (bacteria and archaea) and viruses - to generate the macromolecular machinery for life **[31]**. Scientists study the type and the quantity of mRNAs which are produced by one cell in order to learn which gene are expressed, fact that gives information on how a gene is responding to its needs. Gene expression is a very complex and strictly controlled process that allows to a cell to respond dynamically to its environmental needs. This mechanism performs as a switch on/off in order to control which genes will be expressed in the cell and whether the level of expression of certain genes needs to be increased or decreased.

3.1.3 Analysis of Gene Expression

In genetics, gene expression is the most fundamental level at which the genotype gives rise to the phenotype. The genetic code stored in DNA is "interpreted" by gene expression, and the properties of the expression give rise to the organism's phenotype. Until 1990, scientists could only study a few genes each time. However, nowadays, the use of genetic mechanics made huge steps in the fields of genetic science. The new tool developed the last years is called microarray chip (Figure 3.1) and is known as DNA chip and promises to transfer the science of understanding genes in a new level with the expression of thousands of genes fast and accurately.

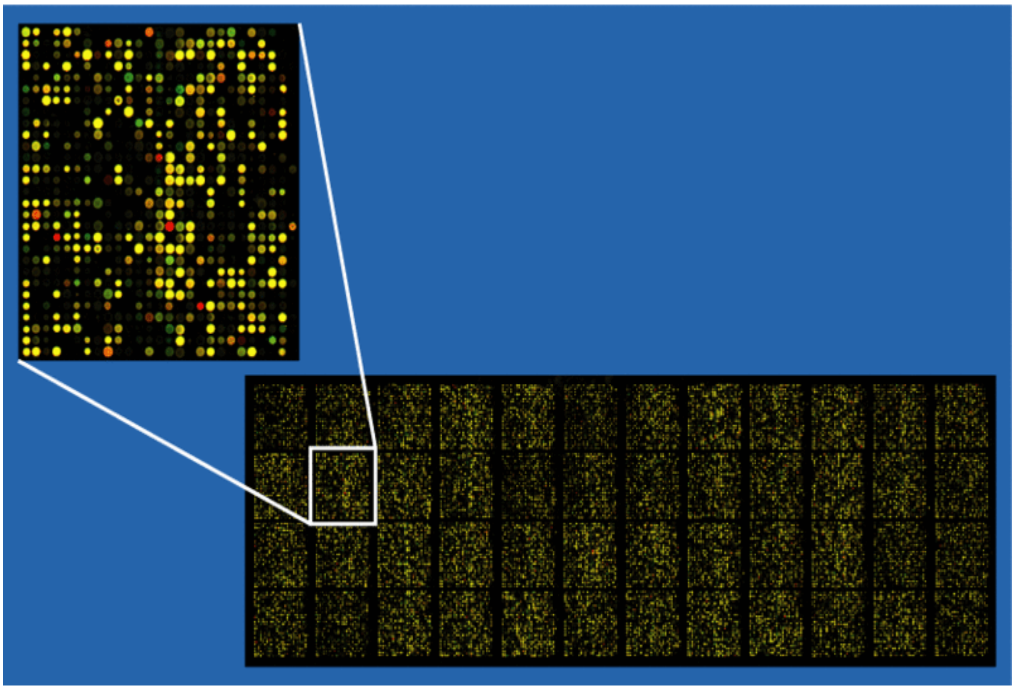


Figure 3.1 – Microarray Chip

3.1.4 Microarrays – How do chips work

The principal behind the analysis of gene expression is based on comparison of samples, for instance tissues – old and new ones -, as well as for the study of the development of healthy and unhealthy tissues of simple and more complex organisms. Figure 3.2 presents the creation of a microarray chip.

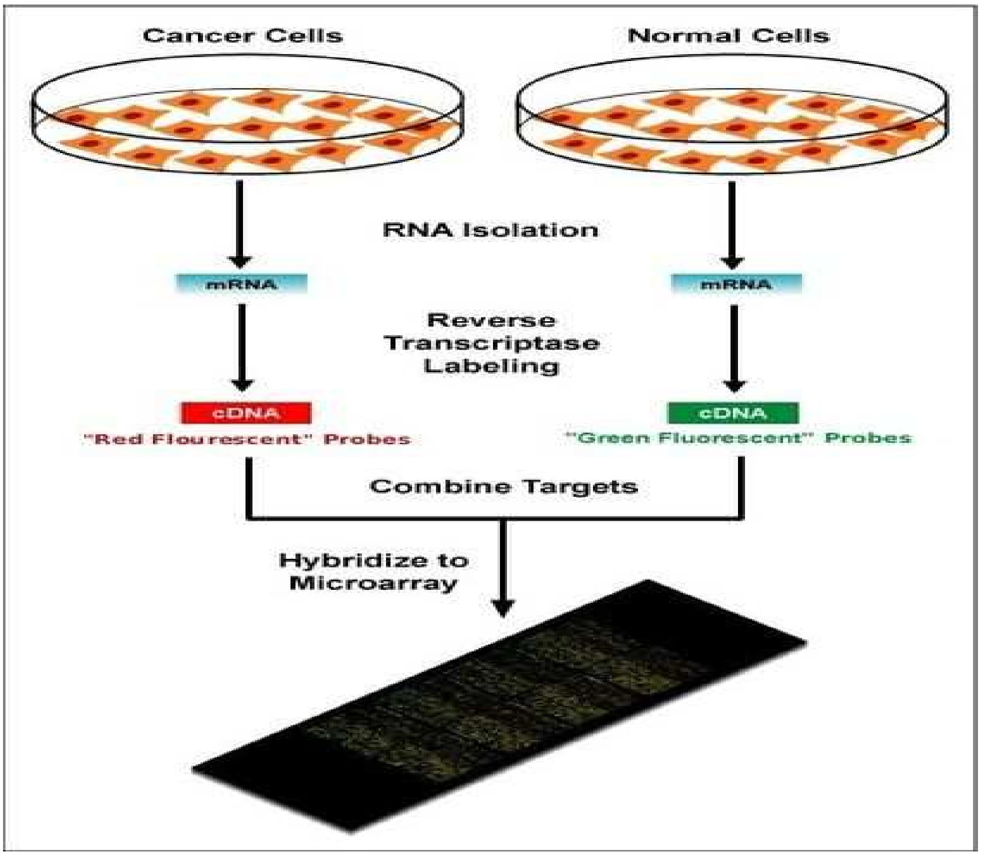


Figure 3.2 – Creation of Microarray Chip

When a gene is expressed in a cell, it generates messenger RNA (mRNA). Over expressed genes generate more mRNA than under expressed genes. This can be detected on the microarray. The first step in using a microarray is to collect healthy and cancerous tissue samples from the patient. This way, doctors can look at what genes are turned on and off in the healthy cells compared to the cancerous cells. Once the tissues samples are obtained, the messenger RNA (mRNA) is isolated from the samples. The mRNA is color-coded with fluorescent tags and used to make a DNA copy (the mRNA from the healthy cells is dyed green; the mRNA from the abnormal cells is dyed red.)

The DNA copy that is made, called complementary DNA (cDNA), is then applied to the microarray. The cDNA binds to complementary base pairs in each of the spots on the array, a process known as hybridization. Based on how the DNA binds together, each spot will appear red, green, or yellow (a combination of red and green) when scanned with a laser.

- A red spot indicates that that gene was strongly expressed in cancer cells.

- A green spot indicates that that gene was strongly repressed in cancer cells.

- If a spot turns yellow, it means that that gene was neither strongly expressed nor strongly repressed in cancer cells.

- A black spot indicates that none of the patient's cDNA has bonded to the DNA in the gene located in that spot. This indicates that the gene is inactive.

3.2 Characteristics of Microarray Data

Studies that are usually done in genetics data use microarray experiments so that parallel comparison between the expressional behaviors of the genes can be made. A gene expression data set from a microarray experiment can be represented by a real valued expression matrix where the rows form the expression patterns of genes, the columns represent the expression profiles of samples, and each cell is the measured expression level of gene *i* in sample *j*[32]. Table 3.1 shows a gene expression matrix.

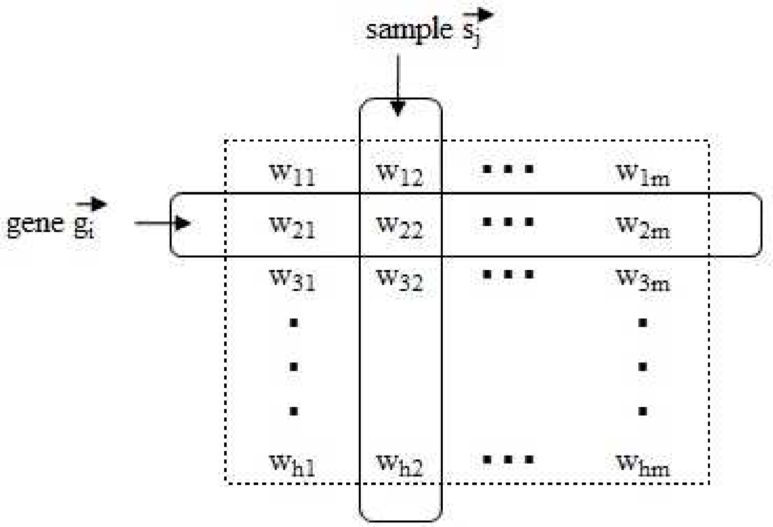


Table 4.1

In general, microarray data have the following characteristics [33]:

- **Dimensionality**: the number of rows (genes) of the matrix can contain thousands genes, while the dimension of the columns (samples), is so much smaller. The cost for a microarray chip limits the number of experiments in a chip.

- **Noise**: in a cDNA microarray experiment, the measurement gene expression level depends on the RNA extraction from a biological sample, the preparation of fluorescently labeled complementary DNA (cDNA) to the corresponding spot on the chip, and the image processing procedure to read out the hybridization intensity. Each of these steps can introduce a considerable amount of noise into the final microarray data matrix.

- **Redundancy**: The biological process under scrutiny in a microarray study is assumably a complicated process, which involves concerted gene reactions in different pathways. While some genes can even be involved in more than one pathway, some others, however, might not be relevant to the biological process. These genes usually show little variation over the different experiments under study. Genes that show little variation over the different experiments are called constitutive with respect to the biological process studied. Constitutive genes often contribute to a large proportion of the whole population of the genes included in a microarray study.

Some problems of data preprocessing have become themselves an interesting research topic. Therefore, some actions must be taken before the analysis of such data.

- Microarray data usually contain **missing values**. The inability of clustering algorithms to face such situation necessitates in the replacement of such values. Most of the times the replacement is done with 0 or with the average of the values. However, such methods can conclude to different clustering results.

- Most of the times it is necessary to normalize data. In microarray data, many noise sources cause systematic sources of biases. A step of **normalization** may help to compute and remove the biases to correct the data.

- After the normalization it is usual to pass the values of genes into a non linear transformation. This method fits in data with ratios of gene expression due to the fact that such ratios are not symmetrical.

3.3 Multivariate techniques in genetics.

The clustering techniques have been proven to be useful to understand gene functions, cells' functions and subcategories of cells. Co expressed genes can be classified with other cell functions. At the same time genes with similar form at the same cluster is possible to be combined with same cell functions.

One of the characteristics of gene expression data is that it makes sense to cluster both genes and the samples. From one side, genes can be clustered in groups based on patterns that they form. On the other hand, samples can be divided in homogeneous groups each one of them will correspond to a particular phenotype, for example a type of cancer.

Clustering techniques can be grouped in 3 forms: class comparison, class discovery and class prediction. In class comparison we observe differences in a constant number of groups and we examine the genes that cause the discrimination. In class discovery we observe groups and patterns in genes.

Finally, in class prediction, we predict the phenotype using the information from gene expression.

3.4 Experimental Results

One of the most important roles of the research of cancer is the development of an accurate classification of cancer cells and cancer tissues. In microarray studies, cluster analysis helps to identify gene groups as well as sample groups. However, what is also important is to define if those clusters are accurate and reproducible as well as biologically significant.

In the following analysis we have used three clustering algorithms, AGNES, DIANA and PAM that were described in the previous chapter. At the same time the algorithms were ran using possible combinations between metric and linkage methods. The combinations are presented in the Table Table 3.2.

Metric	Linkage Method	Notation
Euclidean	Average	*IndexName11*
	Single	*IndexName12*
	Complete	*IndexName13*
	Ward	*IndexName14*
	Weighted	*IndexName15*
Manhattan	Average	*IndexName21*
	Single	*IndexName22*
	Complete	*IndexName23*
	Ward	*IndexName24*
	Weighted	*IndexName25*

Table 3.2 – Combinations of Methods used

Divisive and Partitional Algorithms were run using both metric, the Euclidean and Manhattan Distance. The validation indices used were:

- Davies-Bouldin Index
- Silhouette Index
- Dunn Index

3.4.1. Data

The data used in this analysis are DNA microarray data on primary breast tumors of 78 young patients **[35]**. We applied supervised classification to identify a gene expression signature strongly predictive of a short interval to distant metastases. At the same time, we checked if those clusters appearing were reproducible, with the methods described above.

We selected 78 primary breast cancers: 34 from patients who developed distant metastases within 5 years and 44 from patients who continued to be disease – free after a period of at least 5 years. All patients were under 55 years of age at the time of diagnosis.

From each patient, 5μg total RNA was isolated from snap – frozen tumor material and used to derive complementary RNA (cRNA). A reference cRNA pool was made by pooling equal amounts of cRNA from each of the sporadic carcinomas. Two hybridizations were carried out of each tumor using a fluorescent dye reversal technique on microarrays containing 24,481 human genes synthesized by inkjet technology [34].

The criteria for the sporadic patients were primary invasive breast carcinoma less than 5cm (T1 or T2), no auxiliary metastases, age of diagnosis less than 55 years, calendar year of diagnosis 1983 – 1996, no previous malignancies; all patients were treated by modified radical mastectomy or breast-conserving treatment, including axillary lymph node dissection followed by radiotherapy. Five patients of the metastases group received adjuvant systemic therapy consisting of chemotherapy or hormonal therapy; all other patients did not receive additional treatment. All patients were followed at least annually for a period of at least 5 years. The criteria for hereditary patients were: carriers of a germline mutation in BRCA1 or BRCA2, and primary invasive breast carcinoma; no other selection criterion was applied. This study was approved by the Medical Ethical Committee of the Netherlands Cancer Institute [35].

To gain insight into the genes of the dominant expression signatures, we associated them with hystopathological data; oestrogen receptor (ER) – α expression. We then selected only the genes that their difference in ER receptor was statistically significant.

In order to make all the calculations we used R packages for cluster analysis as well as packages for cluster validations from **R project** [36]. The packages that were used for the analysis were the following: Cluster, ClValid, ClusterSim, ClusterCons. Also, Bioconductor software for R was used which provides tools for the analysis and comprehension of genomic data. Missing values were replaced by the mean of the variable.

3.4.2 Clustering Algorithms Results

An unsupervised, hierarchical clustering algorithm allowed us to cluster the tumors on the basis of their similarities measured over these significant genes.

In Diagram 3.1 we can see a dendrogram produced by the clustering technique. The length and the subdivision of the branches display the relatedness of the breast tumor (right) and the expression of the genes (top).

We can clearly observe two distinct groups of tumors that are the dominant feature in this two dimensional display, suggesting that the tumors can be divided into two types on the basis of this set. Notably in the upper group the patients were from group that developed distant metastasis within 5 years while the lower group shows the patients that were healthy until this time of the analysis.

In general we can see two clusters in genes, that show the distinguish between genes with ERP receptor and genes without. Samples, on the other hand, might also show three clusters but this might be due to the fact that there are some outliers as we can see, for example sample 54. Thus, using unsupervised clustering we can already, to some extent, distinguish between good prognosis and poor prognosis tumors.

Now that we have seen a first clustering of the samples as well as the genes, we can identify if this clustering is reproducible in a different sample. In Diagram 3.2 we selected only the patients that showed metastasis within 5 years while in Diagram 3.3 we only selected patients that did not show any metastasis until that time. Unfortunately the statistically significant genes were only so few and the results were not accurate.

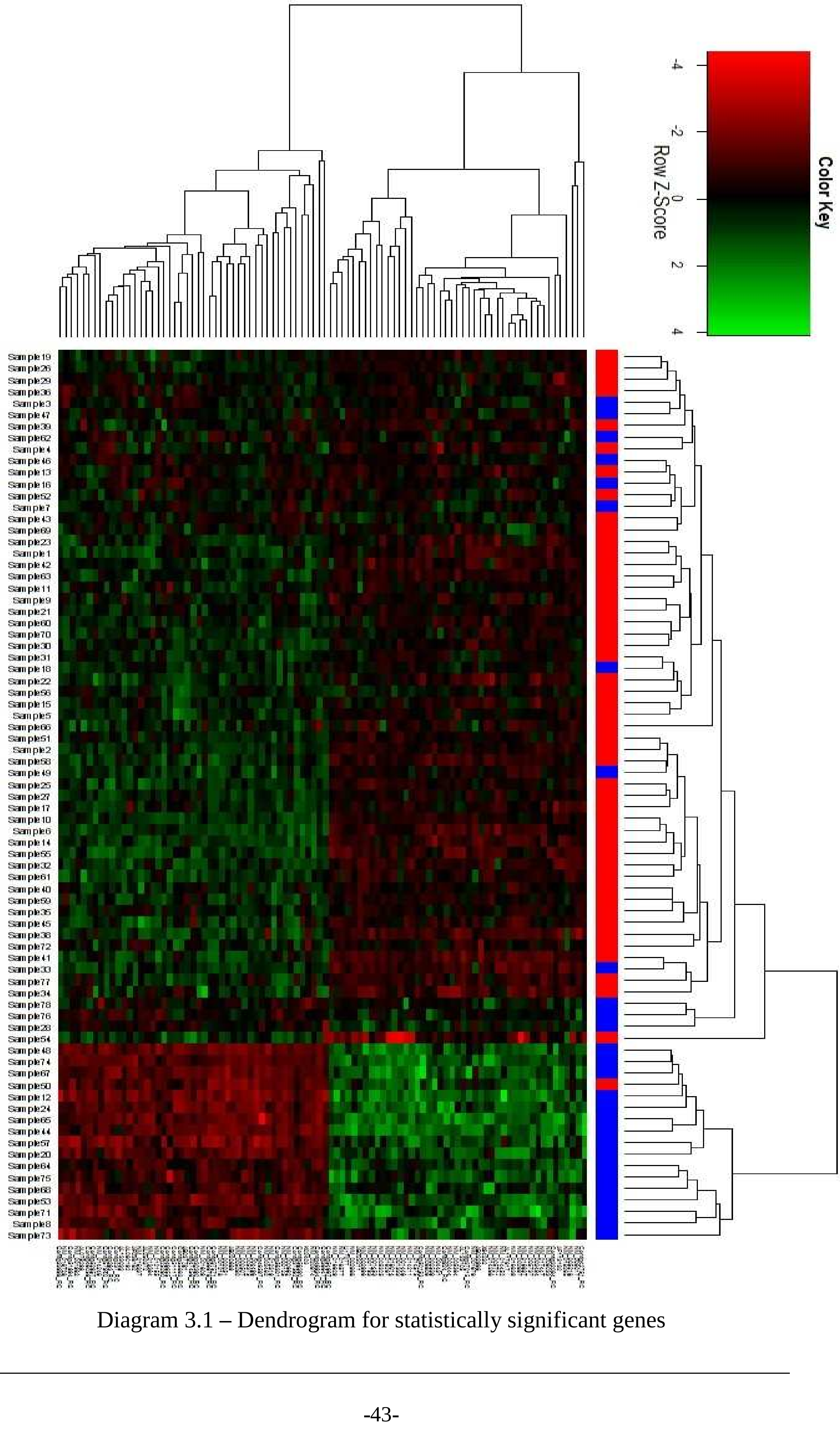


Diagram 3.1 – Dendrogram for statistically significant genes

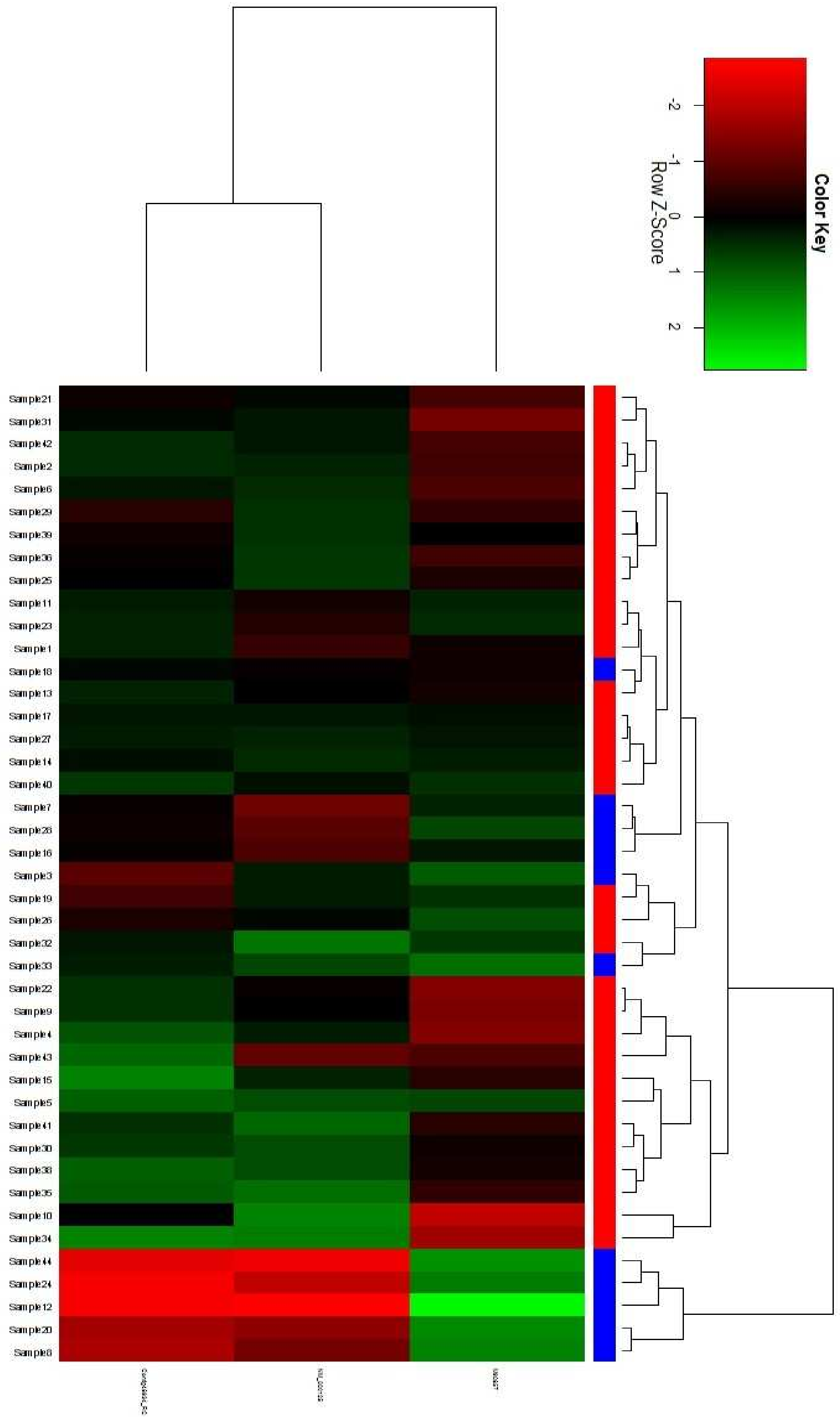


Diagram 3.2 – Dendrogram for patients with metastasis within 5 years

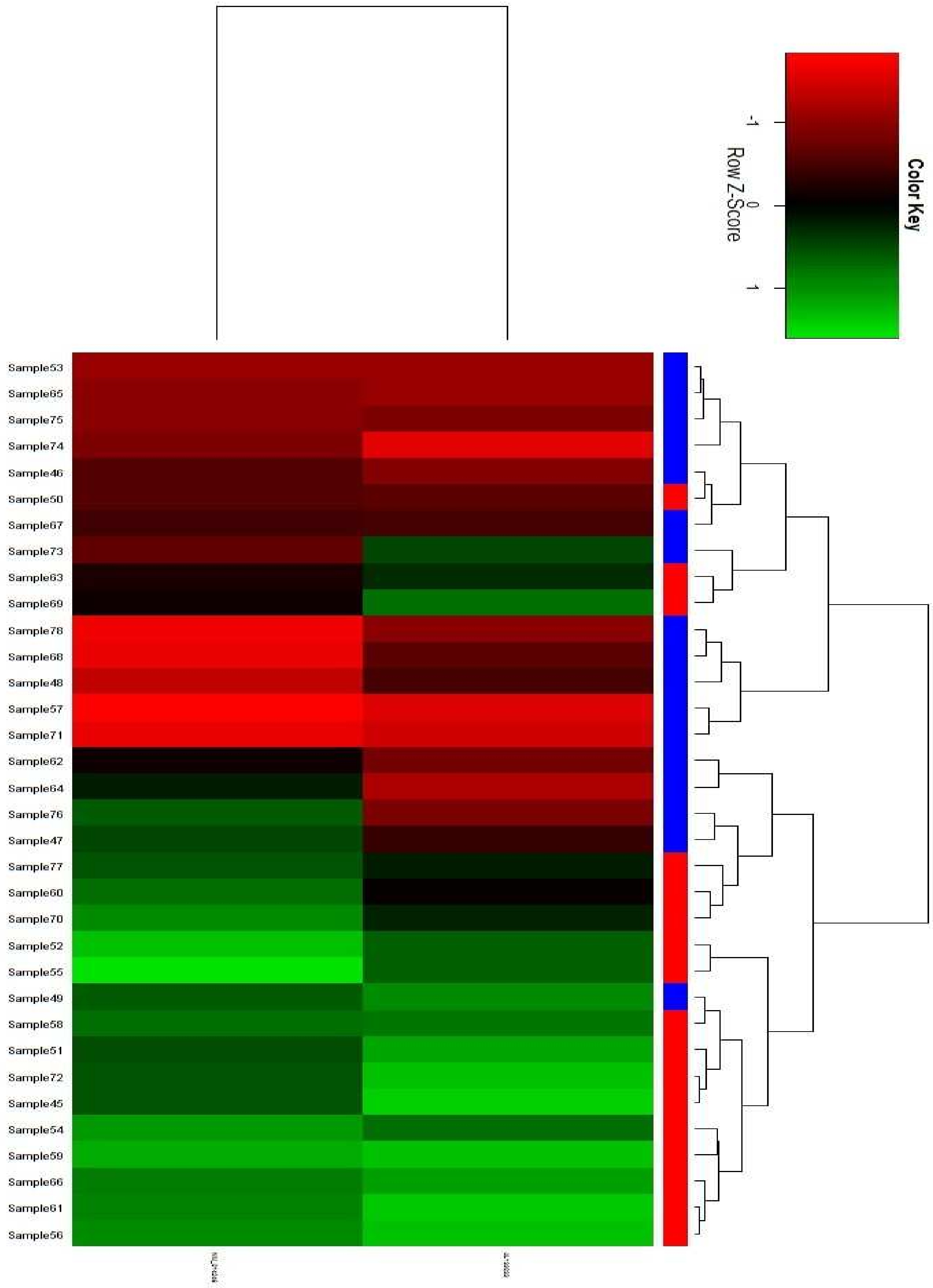


Diagram 3.3 – Dendrogram for patients without metastasis

At a second step we ran some possible clustering algorithms to check if we have different results. Below, we present some of those and discuss about their results.

The Diagram 3.4 shows the hierarchical clustering based on **Agnes Algorithm, manhattan distance and average method**. We can clearly observe that there are two groups in samples as we noticed before and a possible third group appears. However, still we can see that sample 54 shows a different behaviour from the rest.

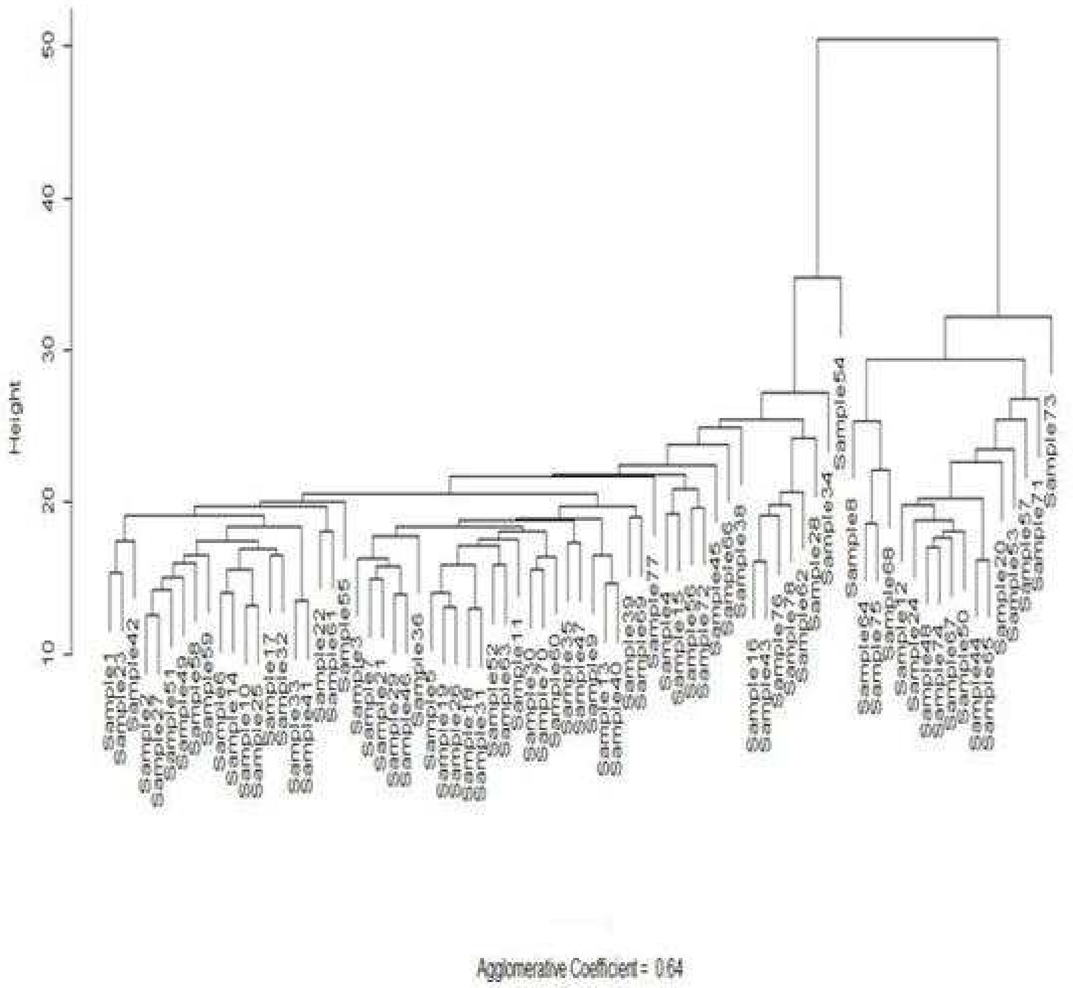


Diagram 3.4 – Dendrogram for AGNES algorithm

If instead of manhattan distance we use euclidean then the results change only slightly.

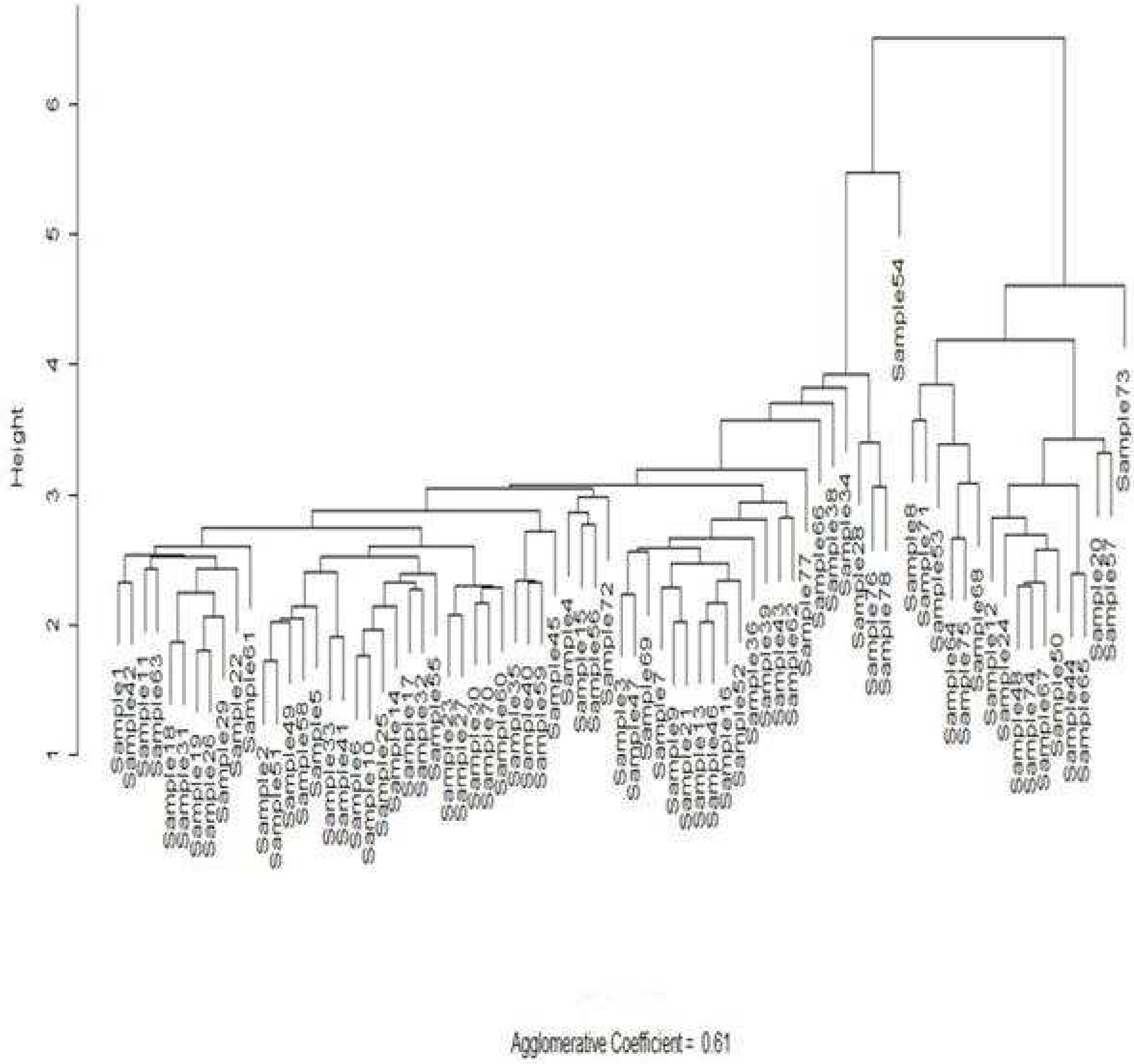


Diagram 3.5 – Dendrogram for AGNES algorithm

The results change when we use the DIANA algorithm which creates more clearly three clusters instead of two. Diagrams 3.6 and 3.7 show the clustering of **DIANA algorithm using euclidean and manhattan distance respectively**.

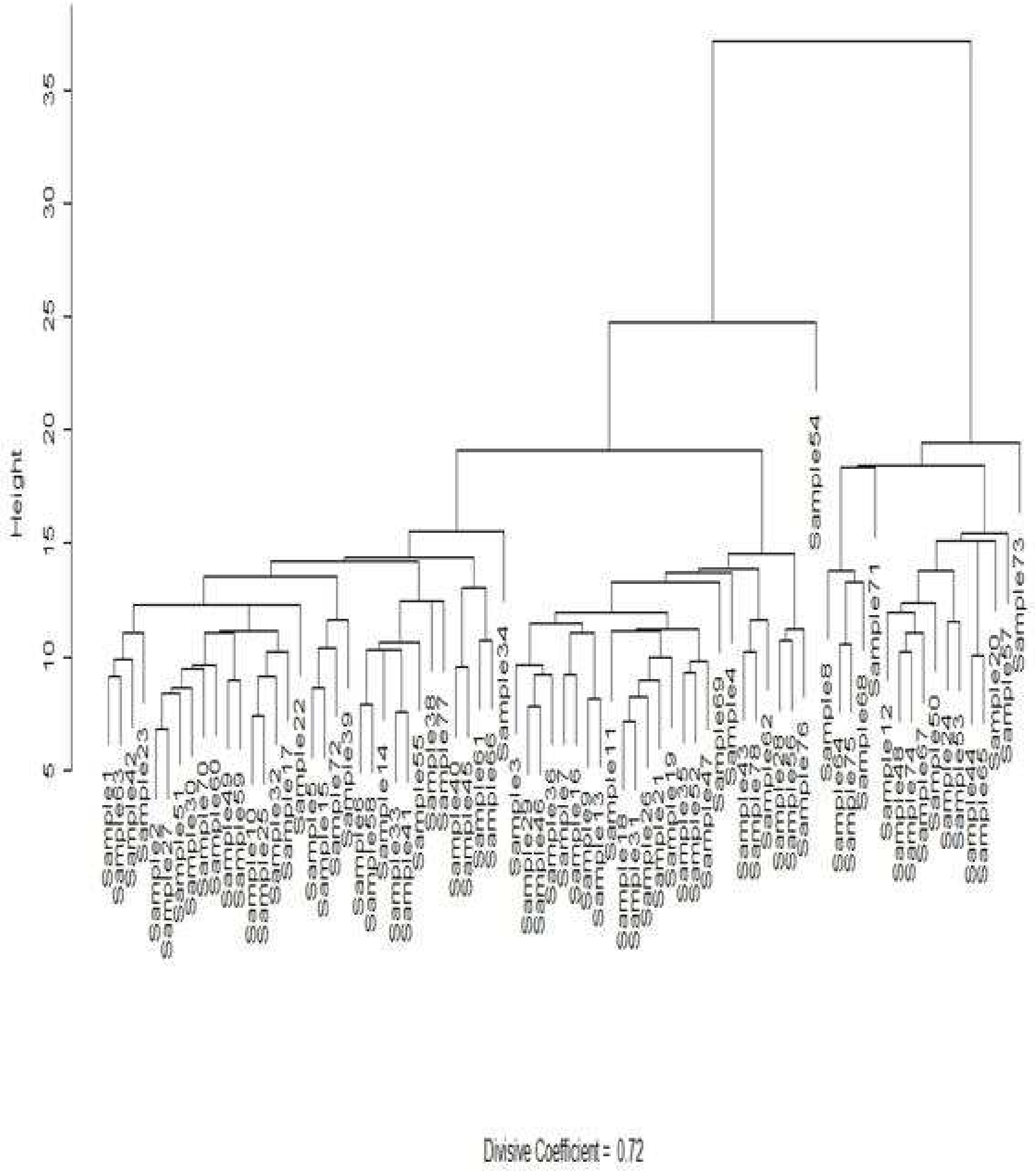


Diagram 3.6 – Dendrogram for DIANA algorithm

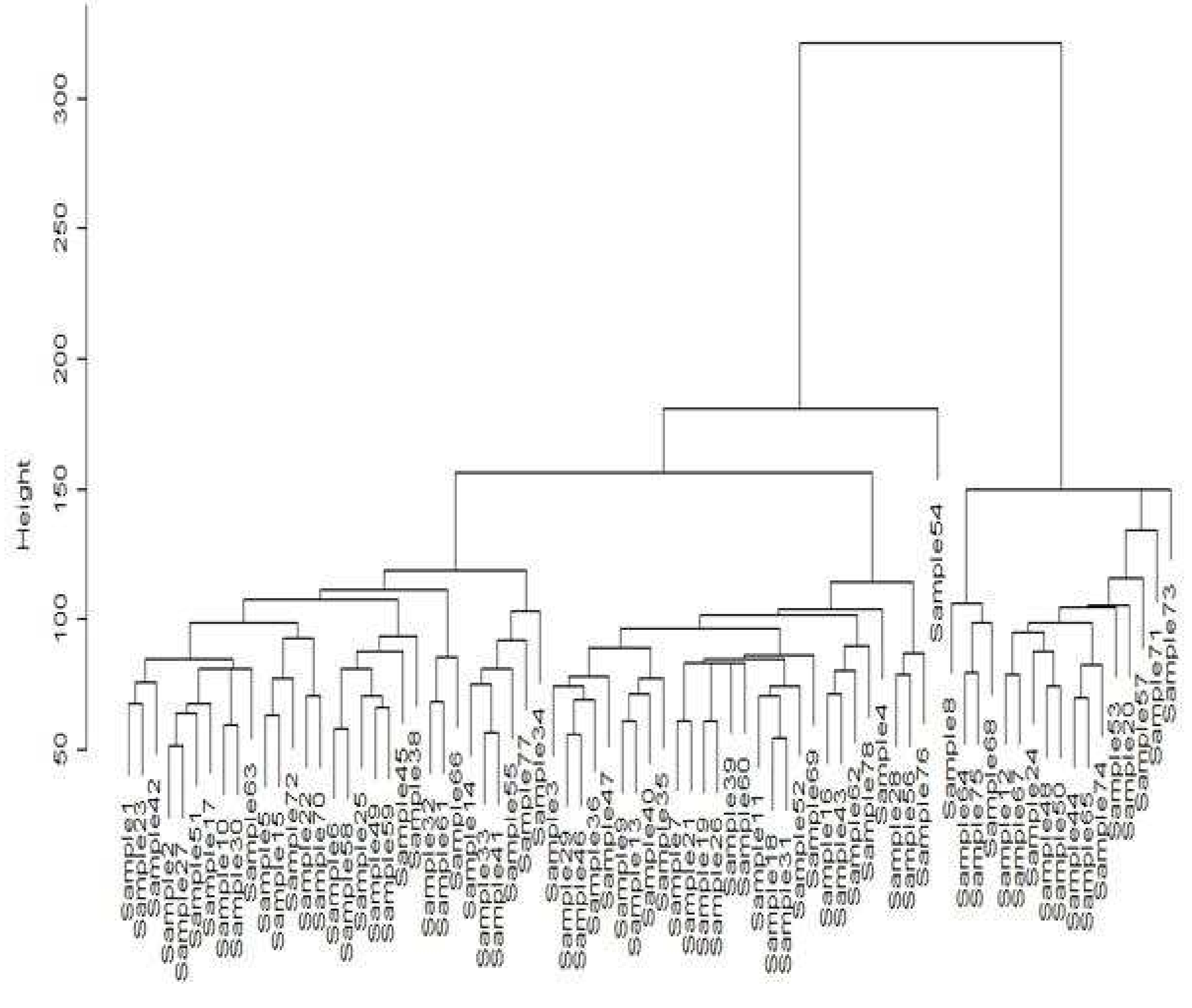


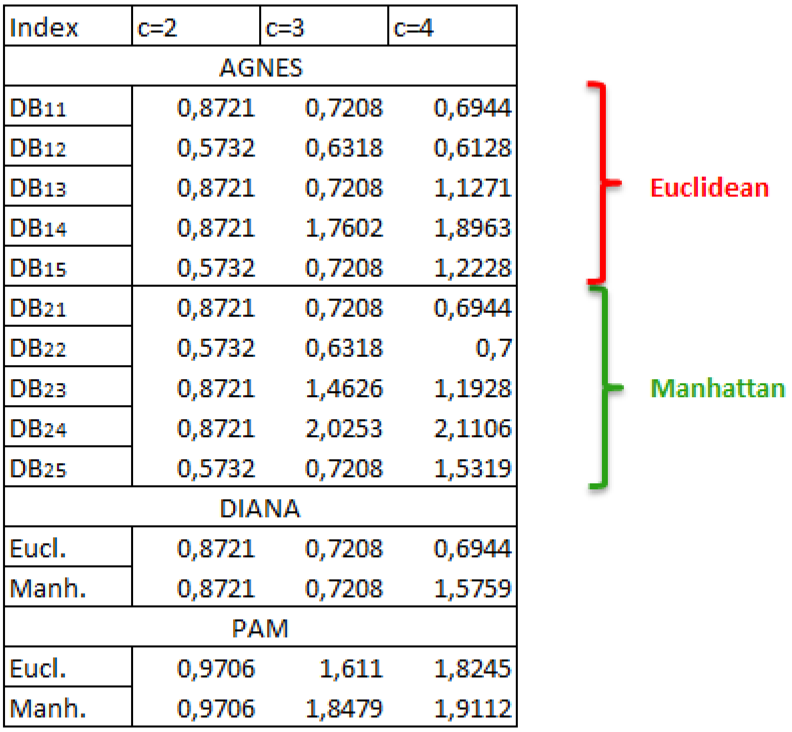
Diagram 3.7 – Dendrogram for DIANA algorithm

3.4.3. Internal Criteria Results

As we have already showed a visible pattern in the data it is necessary to understand if this pattern is reproducible in other datasets and which of the algorithms show more accurate results.

The Davies – Bouldin Index measures how compact and well – separated the clusters are. To obtain clusters with these characteristics the dispersion measure for each cluster needs to be as small as possible, while the dissimilarity measure between clusters needs to be large. According to this, Davies Bouldin index would have small values if the clusters are compact and well separated. In several occasions zero values are obtained. This happens when the clustering algorithm assign one object to each cluster, except in one. That is if the data set consisting of n objects will be divided in three clusters, then two of them will contain only one observation, and one cluster with n-2 observations. Then this zero values are not going to be considerate as a minimum value, because having one object by cluster is not a good clustering result.

Table 4.3 shows the DB index obtained from the data set. The minimum value among all the combinations using AGNES is obtained using Euclidean distance combined with Single method (DB_{12}) for c = 2 clusters. Using DIANA the minimum occurs using Euclidean again for c = 4. Lastly for PAM the minimum occurs using Euclidean for c = 2 clusters. Therefore we conclude that Davies Bouldin index indicates that the best results occur when we have two clusters, as we already saw in the diagrams before.

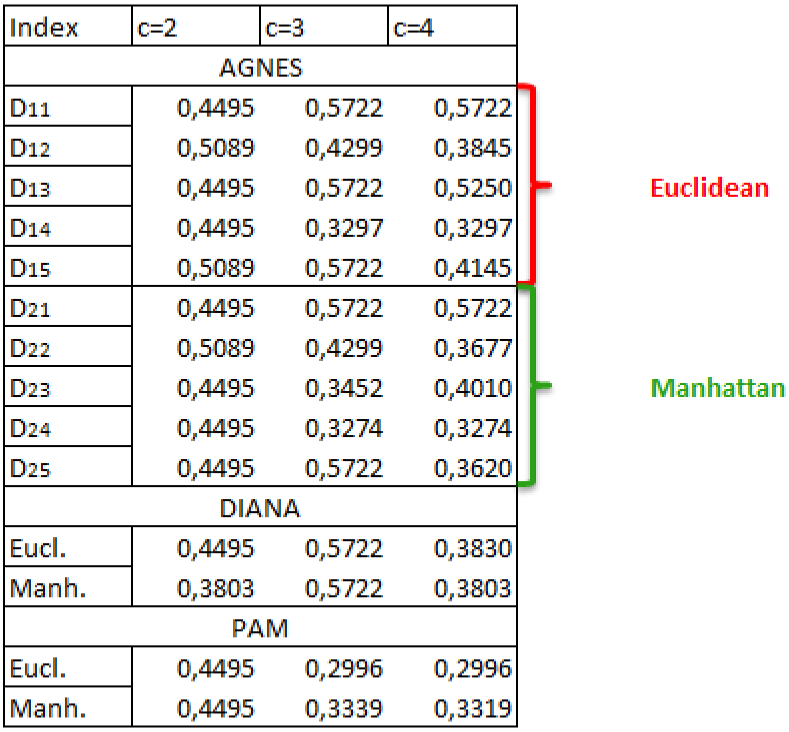

Euclidean
Manhattan

Index	c=2	c=3	c=4
AGNES			
DB$_{11}$	0,8721	0,7208	0,6944
DB$_{12}$	0,5732	0,6318	0,6128
DB$_{13}$	0,8721	0,7208	1,1271
DB$_{14}$	0,8721	1,7602	1,8963
DB$_{15}$	0,5732	0,7208	1,2228
DB$_{21}$	0,8721	0,7208	0,6944
DB$_{22}$	0,5732	0,6318	0,7
DB$_{23}$	0,8721	1,4626	1,1928
DB$_{24}$	0,8721	2,0253	2,1106
DB$_{25}$	0,5732	0,7208	1,5319
DIANA			
Eucl.	0,8721	0,7208	0,6944
Manh.	0,8721	0,7208	1,5759
PAM			
Eucl.	0,9706	1,611	1,8245
Manh.	0,9706	1,8479	1,9112

Table 3.3 – Values of DB index for combinations of algorithms and methods

Opposite to Davies- Bouldin index, the Dunn Index would have large values of the clusters are compact and well-separated.

Dunn Index is presented in Table 3.4. The results here are not so similar. The maximum value among all the combinations using AGNES is obtained using Euclidean distance combined with Average method (DB_{11}) and Complete method (DB_{13}) and Weighted method (DB_{15}) for c = 3 clusters. The same results appear when using Manhattan for c=3 clusters. Using DIANA the maximum occurs using Euclidean or Manhattan again for c = 3. Lastly for PAM the maximum occurs using Euclidean or Manhattan for c = 2 clusters. Therefore we conclude that Dunn index indicates mostly that the best results occur when we have three clusters.

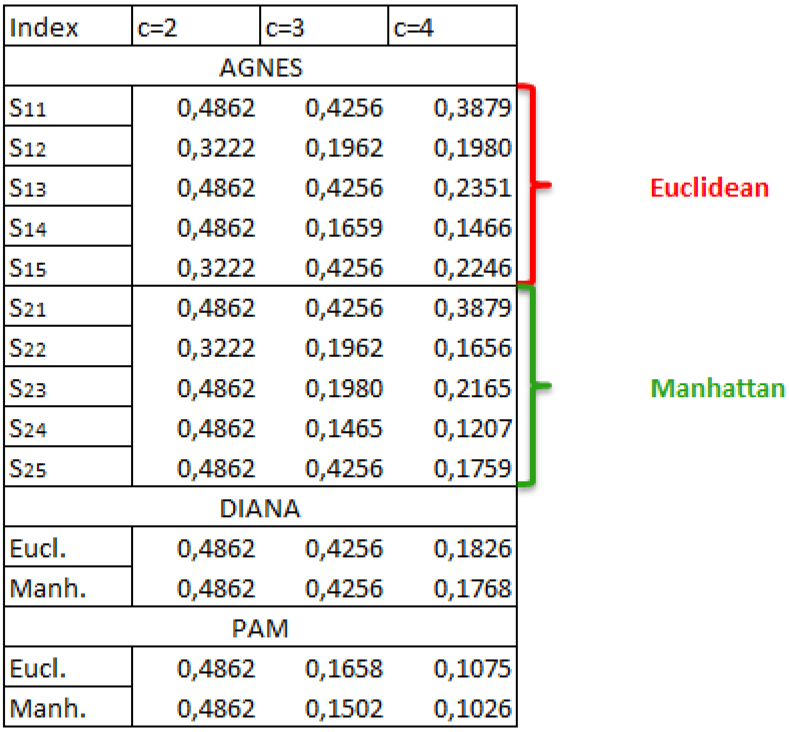


Index	c=2	c=3	c=4	
AGNES				
D11	0,4495	0,5722	0,5722	Euclidean
D12	0,5089	0,4299	0,3845	
D13	0,4495	0,5722	0,5250	
D14	0,4495	0,3297	0,3297	
D15	0,5089	0,5722	0,4145	
D21	0,4495	0,5722	0,5722	Manhattan
D22	0,5089	0,4299	0,3677	
D23	0,4495	0,3452	0,4010	
D24	0,4495	0,3274	0,3274	
D25	0,4495	0,5722	0,3620	
DIANA				
Eucl.	0,4495	0,5722	0,3830	
Manh.	0,3803	0,5722	0,3803	
PAM				
Eucl.	0,4495	0,2996	0,2996	
Manh.	0,4495	0,3339	0,3319	

Table 3.4 – Values of Dunn Index for combinations and methods

Finally, the best results for clustering occur when we have maximum value of Silhouette index (Table 3.5). The maximum value for AGNES occurs when we have Euclidean or Manhattan distance with almost all methods for c = 2 clusters. For DIANA maximum value occurs again for c = 2 clusters using both distances. Lastly for PAM again c = 2 clusters maximize the Silhouette Index. It is clear that for Silhouette index the most well separated results are for two clusters.

In general the best results seem to occur when we have two clusters as we also observed from the diagrams.


Euclidean
Manhattan

Index	c=2	c=3	c=4
AGNES			
S_{11}	0,4862	0,4256	0,3879
S_{12}	0,3222	0,1962	0,1980
S_{13}	0,4862	0,4256	0,2351
S_{14}	0,4862	0,1659	0,1466
S_{15}	0,3222	0,4256	0,2246
S_{21}	0,4862	0,4256	0,3879
S_{22}	0,3222	0,1962	0,1656
S_{23}	0,4862	0,1980	0,2165
S_{24}	0,4862	0,1465	0,1207
S_{25}	0,4862	0,4256	0,1759
DIANA			
Eucl.	0,4862	0,4256	0,1826
Manh.	0,4862	0,4256	0,1768
PAM			
Eucl.	0,4862	0,1658	0,1075
Manh.	0,4862	0,1502	0,1026

Table 3.5 – Values of Silhouette Index for combinations and methods

From the three measures that we used in this analysis, the best one appears to be AGNES. For Silhouette Index and Dunn Index, it gives the highest values, while for Davies – Bouldin index it gives the smallest values.

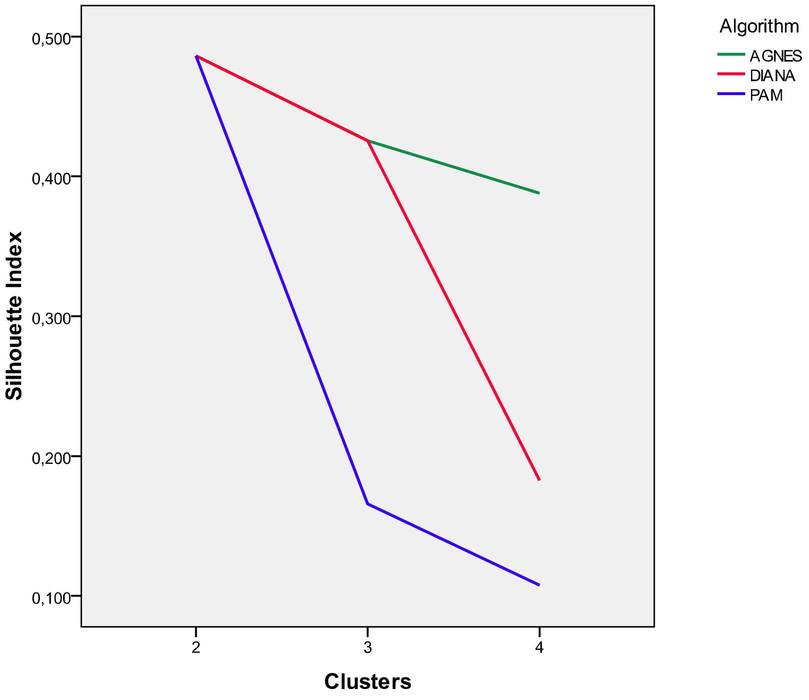


Diagram 3.8 – Diagram for Values of Silhouette Index

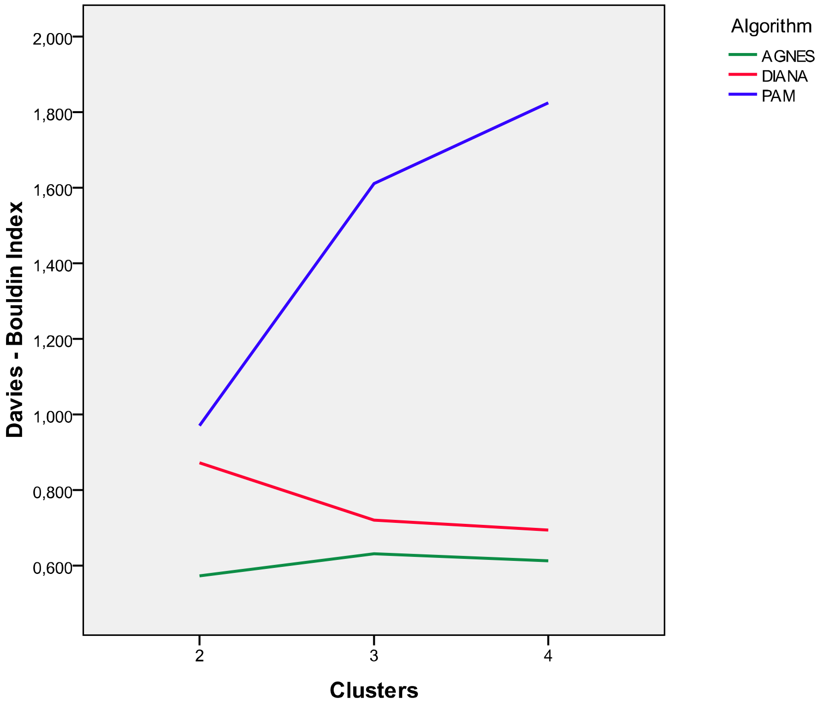


Diagram 3.9 – Diagram for Values of DB Index

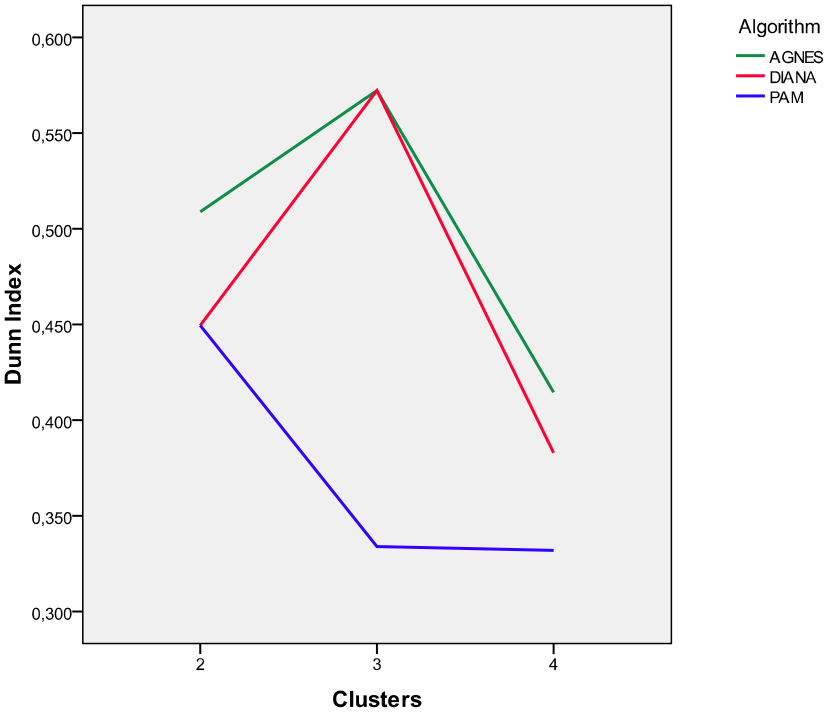


Diagram 3.10 – Diagram for Values of Dunn Index

Chapter 4

4. Conclusion

In this thesis the main objective was to compare some of the validation indices in order to detect the optimal number of classes a data set can have.

The best results were obtained when using Complete and Ward Linkage Methods. In general PAM clustering results were not good. On the other hand, when using DIANA the results were similar to the ones obtained with AGNES, and there were no significant difference between using Euclidean or Manhattan Distance. In all cases, AGNES seems to do a better clustering task.

In future studies we can include more validation indices as well as external validation metrics. This can be conducted with the inclusion of data from other sources so that we can validate our clustering results externally. Data sources such as genomic data obtained in molecular biology labs and family information collected from the siblings of the patients in this study can be used for the purpose of external validation.

Chapter 5 - Bibliography

1. Broke, G., Pihur, V., Datta, S. and Datta, S., ‘clValid : An R Package for Cluster Validation’, *Journal of Statistical Software*, vol. 25, no. 4, 2008.
2. Καρλής , Δ. , *Πολυμεταβλητή Στατιστική Ανάλυση*, Εκδόσεις Σταμούλη Α.Ε., Αθήνα, 2005.
3. Theodoridis, S. and Koutroumbas, K., *Pattern Recognition*, 3rd edn, Elsevier, London, 2006.
4. Mahalanobis, P C. , *On the generalised distance in statistics* . Proceedings of the National Institute of Sciences of India 2 (1): 49–55, 1936
5. Kruskal, J.B., *Multidimensional scaling by optimizing goodness of fit to a non metric hypothesis*. Psychometrika 29(1):1-27, 1964
6. Kaufman, L., and Rousseeuw, P., *Finding groups in data: An introduction to cluster* analysis. John Wiley & Sons, 1990.
7. Theodoridis, S., and Koutroumbas, K., *Pattern Recognition*. Academic Press, 1999.
8. Belitskaya – Lévy, I. , *Elements of Applied Statistics and Data Mining, with Applications to Biology and Medicine*, 2004, http://www.med.nyu.edu/biostatistics/people/Ilana%20Belitskaya-Levy/Courses/MAS/Handouts/hier.pdf
9. Raymond, T., Jiawei, H. , *Efficient and effective clustering methods for spatial data mining*, VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases, 1994.
10. Hartigan, J., and Wong, M., *A k-means clustering algorithm*. Journal of applied statistics, 28. 1979.
11. Batistakis, Y., Halkidi, M., and Varzigiannis, M., *On clustering validation techniques*. Journal of Intelligent Information System, 17:2/3, 2001.
12. Kovács, F., Legány, C., and Babos, A., *Cluster Validity Measurement Techniques*, the 6th International Symposium of Hungarian Researchers on Computational Intelligence, Budapest, 2005

13. Handl, J., Knowles, J., and Kerr, D., *Computational cluster validation in post-genomic data analysis*. Bionformatics Journal, Vol.21 no15, pp 3201-3212, 2005.

14. Acar, A.C., *Cluster Validation* Data Mining Course Bilkent University, 2010, http://www.cs.bilkent.edu.tr/~aacar/courses/CS558/F10/ClusterValidation.pdf.

15. Maimon, O., Rokach, L., *The Data Mining and knowledge Discovery handbook*, 1st end, Springer Publisher, 2005.

16. Mitra, B. and Ho, T.K., *Data Complexity in Pattern Recognition*, 1st edn, Springer Publisher, London, 2006.

17. Jiang, X. and Petkov, N., *Computer Analysis of Images and Patterns*, 13th International Conference CAIP, Munich, 2009.

18. Alippi, C., Polykarpou, M., Panayiotou, C. and Ellinas, G., *Artificial Neural Networks ICANN 2009*, 19th International Conference Limassol, 2009.

19. Fowlkes, E., and Mallows, C. , *Method for comparing two hierarchical clusterings*. Journal of the American Statistical Asociation, 78, 1983.

20. Yeung, K., Ruzzo, W., *An empirical study on Principal Component Analysis for clustering gene expression data*, Technical Report UW-CSE-2000-11-03, 2000

21. Pacual D., Pla F., Sanchez J.S. *Cluster validation using information stability measures*, Pattern Recognition Letters 31, 2010, pp.454-461.

22. Batistakis Y., Halkidi, M. and M. Vazirgiannis. *Clustering validity cheking: Part ii*. Sigmod Record, 31(3), 2002.

23. Bolshakova, N., and Azuaje, F., *Cluster validation techniques for genome expression data*. Signal Processing, 83:825–833, 2003.

24. Rousseuw, P., *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*. Computational and Applied Mathematics, 20, 1987.

25. http://www.mathworks.com/help/toolbox/stats/bq_679x-18.html

26. [Ashis, S., *Advances in Multivariate Statistical Methods*, World Scientific Publishing Co, 2009

27. Petrović, S., *A Comparison Between the Silhouette Index and the Davies-Bouldin Index in Labelling IDS Clusters*, Proceedings of the 11th Nordic

Workshop on Secure IT-systems, NORDSEC 2006, pp. 53-64, Linkoping, Sweden, 2006.

28. Yang,C. et al., *Arguing the validation of Dunn's Index in Gene Clustering,* Biomedical Engineering and Informatics, 2009.

29. Bolshakova, N., Azuaje, F., *Cluster validation techniques for genome expression data* , Biometrics Journal , Volume 83 Issue 4, April 2003

30. Tibshirani, R. , Kapp, Amy, *Are clusters found in one dateset present in another dataset,* Biostatistics , 8, 1, pp. 9-31, 2007.

31. www.wikipedia.com

32. Jiang, D. , Tang, C. and Zang, A. , *Cluster Analysis for Gene Expression Data: A Survey,* IEEE Transactions of Knowledge and data engineering, vol.16, no 11, 2004.

33. Sheng, Q., Moreau, Y., De Smet, F., Marchal, K. and De Moor, B., *Cluster analysis of microarray data,* Notes from Center for Biological Sequence Analysis Danish Technical University

34. Hughes, T., *et al., Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer.* Nature Biotechnology, 19, 342-347, 2001

35. Van't Veer, L. et al., *Gene expression profiling predicts clinical outcome of breast cancer,* Nature, Vol.415, 2002

36. http://www.r-project.org/