A COMPARISON IN CLUSTER VALIDATION TECHNIQUES

By

Marggie D. González Toledo

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

MATHEMATICS (STATISTICS)

UNIVERSITY OF PUERTO RICO
MAYAGÜEZ CAMPUS

December, 2005

Approved by:

Tokuji Saito, Ph.D Member, Graduate Committee	Date
Edgardo Lorenzo, Ph.D Member, Graduate Committee	Date
Edgar Acuña, Ph.D President, Graduate Committee	Date
Jorge L. Ortiz, Ph.D Representative of Graduate Studies	Date
Pedro Vasquez, D.Sc. Chairperson of the Department	Date

Abstract of Dissertation Presented to the Graduate School
of the University of Puerto Rico in Partial Fulfillment of the
Requirements for the Degree of Master of Science

A COMPARISON IN CLUSTER VALIDATION TECHNIQUES

By

Marggie D. González Toledo

December 2005

Chair: Edgar Acuña, Ph.D.
Major Department: Mathematics Department

Clustering may be defined as a process that aims to find partitions of similar objects. It is an unsupervised recognition procedure since there are no predefined classes that indicate grouping properties in the data set. Researchers have extensively studied clustering since it arise in many application domains in engineering, social science, and biology. The basic problem in clustering is to decide the optimal number of clusters, or partitions, that fits a data set. Sometimes the clusters obtained after we applying some clustering algorithms does not represent the structure that the data set really has. For this reason we need quantitative measures to evaluate the results of a clustering algorithm. This task is named *Cluster Validity*.

This thesis includes a description about the clustering algorithms, and its validation techniques. Our main goal is to identify which cluster validation techniques is most efficient in order to divide a given data set. In this research it was done applying seven cluster validation techniques along with three clustering algorithms on ten different data sets. The results were obtained using the R programming language and environment for statistical computing. This software can be download from the page http://www.r-project.org/ [1].

Resumen de Disertación Presentado a Escuela Graduada
de la Universidad de Puerto Rico como requisito parcial de los
Requerimientos para el grado de Maestría en Ciencias

UNA COMPARACIÓN DE ÍNDICES DE VALIDACIÓN DE CONGLOMERADOS

Por

Marggie D. González Toledo

Diciembre 2005

Consejero: Edgar Acuña, Ph.D.
Departamento: Departamento de Matematicas

Análisis de Conglomerados puede definirse como el proceso que intenta encontrar particiones de objectos similares. Es un procedimiento de reconocimiento no supervisado porque no hay clases predefinidas que indiquen propiedades de agrupamiento en la base de datos. Decidir el número de particiones en los que se debe dividir un conjunto de datos es un problema que hay que enfrentar cuando se trabaja con análisis de conglomerados. En algunas ocasiones los grupos obtenidos después de aplicar algún algorítmo de conglomerados, no representan la estructura real que la base de datos posee. Por esta razón se necesitan medidas cuantitativas para evaluar el resultado del algoritmo de conglomerados. Esta tarea es llamada *Validación de Conglomerados*.

Esta tesis incluye una descripción de los algorítmos de conglomerados, así como de las técnicas de validación. Nuestra meta principal es identificar que técnica de validación de conglomerados es más efectiva cuando se trata de identificar si un conjunto de datos está bien dividido. En esta investigación se aplicaron siete técnicas de validación junto con tres algorítmos de conglomerados en diez bases de datos

diferentes. Los resultados fueron obtenidos usando el lenguaje de programación y ambiente para computación estadística R que puede obtenerse accesando la página electrónica http://www.r-project.org/ [1].

In our life there are people who inspire, people who motives, people who instruct, and people who encourage you to attain your objectives.

I want to dedicate this work to my family... the reason of my inspiration: to my mother Maggie Toledo, because you trust in me and because you taught me the importance of trust in God; to my father Ovidio González, because you taught me to value the things obtained in life; thanks to both of you for all your sacrifices during these years in order to give your sons the necessary things to have a good life. To who offer me motivation, my brothers Ovimael and Joel. I also want to dedicate this work to a person who takes care of me by means of her prayers, my grandmother, Carmen Toledo.

To Pedro Torres... my husband, thanks for all your valuable help during these years, and thanks for encourage me to obtain this degree. I love you.

ACKNOWLEDGMENTS

I would like to thanks to Edgar Acuña, my thesis advisor, because of his support, and his enthusiasm in helping me during this years. Because of his help in the writing of this document.

Thanks to Luis F. Cáceres, Nilsa Toro and Madeline Ramos, because you always have been there and because I know that you always will be there.

To my mates, and friends Victor, Angel, Santiago, Jose, Gerardo, Jhonny, Juan, Milena, Trilce, Sindy y Karen, because of the moments that we has had together.

To Nina, Betzy, and Judith... You are definitely my best friends for ever.

To all my family because of their advices, and their prayers.

To the Department of Mathematics.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure page

CHAPTER 1
INTRODUCTION

Pattern recognition is the discipline in which objects are classified into a number of groups. When it is available training data with a known label class, it is named supervised pattern recognition. Most of the times there is no such training data. In this situation the main goal is to unreveal the similarities, and cluster similar objects together. This is known as unsupervised pattern recognition or clustering. In clustering a major issue is to define the "similarity" between two objects and choose an appropriate measure for it. Another important issue is how to choose the algorithm that will cluster the objects using the similarity measure chosen before. Different algorithm schemes may give different results, and the expert need to interpret it.

Cluster analysis is one of the most important task in the data mining process for discovering groups. Nowadays clustering analysis methods are applied in many domains. Cluster analysis seeks to separate a set of data into groups or clusters in which the objects into each cluster are more similar to each other than objects in other clusters.

The main goal of this thesis is to use the validation index among the clustering algorithm using ten data sets. Once a clustering algorithm has been applied, seven validation index were computed in order to conclude how many cluster does the data set will have. Three of these data set were obtained from the web site of the University of California at Irvine, and the remaining were obtained from several web sites containing microarray data sets [2].

Each data set was classified using three clustering algorithm: Hierarchical Agglomerative Nesting (AGNES), Hierarchical Divisive Analysis (DIANA), and Partitioning Around Medoids (PAM). Two similarity measure, Euclidean distance and Manhattan distance, were used for all the clusters algorithms. In the AGNES algorithms five linkage methods were considered.

CHAPTER 2
CLUSTERING ALGORITHMS

2.1 Introduction

Cluster analysis is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity. Clustering is a special kind of classification imposed on a finite set of objects and may be defined as a process that aims to find partitions groups of similar objects [3]. The relationship between objects is represented in a *Proximity Matrix* (PM), in which rows and columns correspond to objects. If the objects are characterized as points in a d-dimensional metric space, each entrance of the PM represent the distance between pairs of points, such as Euclidean distance.

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 2.2, a mathematical definition for clustering is presented. Then, in Section 2.2.1, similarity and dissimilarity measures, needed to know how similar or dissimilar two objects are, will be presented. Also it is necessary to have measures that tell us how similar or dissimilar two clusters are. Because of this, in Section 2.2.2, the Linkage Methods used to join, or separate clusters are discussed.

Clustering Algorithms considered in this thesis are divided in two groups: (1) Hierarchical and (2) Partitioning Algorithms. These algorithms are discussed in Sections 2.3 and 2.4, respectively.

2.2 Cluster: A Definition

When we talk about clustering, first we need to define what a cluster means. Many definitions have been proposed over the years. Let us use the one that is in [4]. Let X be a data set, that is,

$$X = \{x_i, i = 1, ..., N\}. \tag{2.1}$$

Now, let be the partition, $\Re$, of X into m sets, $C_j, j = 1, ..., m$. These sets are called clusters and need to satisfied the following three conditions:

- $C_i \neq \emptyset, i = 1, ..., m$
- $\bigcup_{i=1}^{m} C_i = X$
- $C_i \cap C_j = \emptyset, i \neq j, i, j = 1, ..., m$

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 the 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 will be discussed in the next section.

2.2.1 Similarity and Dissimilarity Measures

Because the intention in the clustering algorithms is to join (or separate) the most similar (or dissimilar) objects of a data set X, it is necessary to apply a function that can make a quantitative measure among vectors. This quantitative measure can be arranged in a matrix called *Proximity Matrix*. Two types of quantitative measures can be considered: Similarity Measures, and Dissimilarity Measures.

Similarity Measures

Similarity Measures are used to find similar pairs of object among X. Let $s(i,j)$ be a similar coefficient. If objects i and j are alike, then $s(i,j)$ becomes larger. Otherwise, $s(i,j)$ becomes smaller. For all objects i and j, a similarity measure need to satisfy the following conditions:

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

Dissimilarity Measures

Dissimilarity Measures are used to find dissimilar pairs of object among X. The dissimilarity coefficient, $d(i,j)$, are small when objects i and j are alike, otherwise, $d(i,j)$ become larger. As the similarity measures, 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. Two of the measures most used in practice, and the ones used in this thesis are:

- ***Euclidean Distance***

  The *Euclidean Distance* between points $x = (x_1, x_2, .., x_n)$ and $y = (y_1, y_2, .., y_n)$ is given by:

$$d_1(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.2}$$

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

- ***Manhattan Norm***

  The *Manhattan Norm* between points $x = (x_1, x_2, .., x_n)$ and $y = (y_1, y_2, .., y_n)$ is given by:

$$d_2(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{2.3}$$

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

2.2.2 Linkage Methods

The *Linkage methods* are the quantitative measures used to join the two most similar clusters in the agglomerative clustering algorithm [5].

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 objets i, and j where i is an object of C_i and j is an object of C_j. The Linkage Methods considered in this thesis are:

- ***Unweighted pair-group Method using Arithmetic Averages*** (UPGMA), also called *Group Average Method*

  The *UPGMA* was introduced by Sokal and Michener in 1958 [6] [7]. 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:

$$\delta_1(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{\substack{i \epsilon C_i \\ j \epsilon C_j}} d(i, j) \tag{2.4}$$

- ***Single Linkage Clustering Method*** (SLINK)

  In *SLINK*, introduced by Florek et al. in 1951 [7], the distance between two clusters is taken to be the minimum of all the pairwise distances. Then, the SLINK is defined as follows:

$$\delta_2(C_i, C_j) = \min_{\substack{i \epsilon C_i \\ j \epsilon C_j}} d(i, j) \tag{2.5}$$

- ***Complete Linkage Clustering Method*** (CLINK)

  The *CLINK* is exactly the opposite of the SLINK. The CLINK is defined as follows:

$$\delta_3(C_i, C_j) = \max_{\substack{i \epsilon C_i \\ j \epsilon C_j}} d(i, j) \tag{2.6}$$

- ***Ward's Minimum Variance Clustering Method***

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

$$\delta_4^2(C_i, C_j) = \frac{2|C_i||C_j|}{|C_i| + |C_j|} ||\mu_i - \mu_j||^2 \tag{2.7}$$

  where $\mu_i = \frac{1}{|C_i|} \sum_{x \epsilon C_i} x$ and $\mu_j = \frac{1}{|C_j|} \sum_{x \epsilon C_j} x$

- ***Weighted pair-group Method using Arithmetic Averages*** (WPGMA)

  The *WPGMA* was introduced by Sokal and Sneath in 1963 [7]. It 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 C_i and C_j, forming some new cluster C_k, the dissimilarities are updated by,

$$\delta_5(C_k, C_m) = \frac{1}{2} d(C_i, C_m) + \frac{1}{2} d(C_j, C_m) \tag{2.8}$$

  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.

2.3 Hierarchical clustering algorithms

Hierarchical clustering techniques organize the data into a nested sequence of groups. 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. 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 [8] [9].

2.3.1 Hierarchical Agglomerative Nesting Algorithm (AGNES)

This method starts with N cluster, each containing a single object of X. In the next step, the most similar objects are assigned to a small cluster. The process continues until all the objects lie in a single cluster [10] [7].

Let $d(C_i, C_j)$ be a function that measures the proximity between C_i and C_j, and t the current level of hierarchy. Then, the general scheme can be described as follows [4]:

(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 join 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 most 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 most closest clusters as in the previous step.

(e) Repeat until all the vectors lies in a single cluster.

Now, an example is given in order to explain how AGNES works. The Dissimilarity measure used was the Euclidean Distance, and the Linkage Method used was the Unweighted pair-group Method using Arithmetic Averages.

In Table 2–1, a data set consisting of five people with their weights (in Kilograms), and their heights (in centimeter) is given.

Next, an Agglomerative Nesting Hierarchical Clustering will be carried out following, step by step, the algorithm discussed previously.

Before we begin with the algorithm it is necessary to standardized the data set. Table 2–2 contains the standardized data.

Name	*Weight* (kg)	*Height* (cm)
Joshua	15	95
Anne	49	156
Bryan	13	95
Peter	45	160
Cristin	85	178

Table 2–1: Data set, AGNES Example

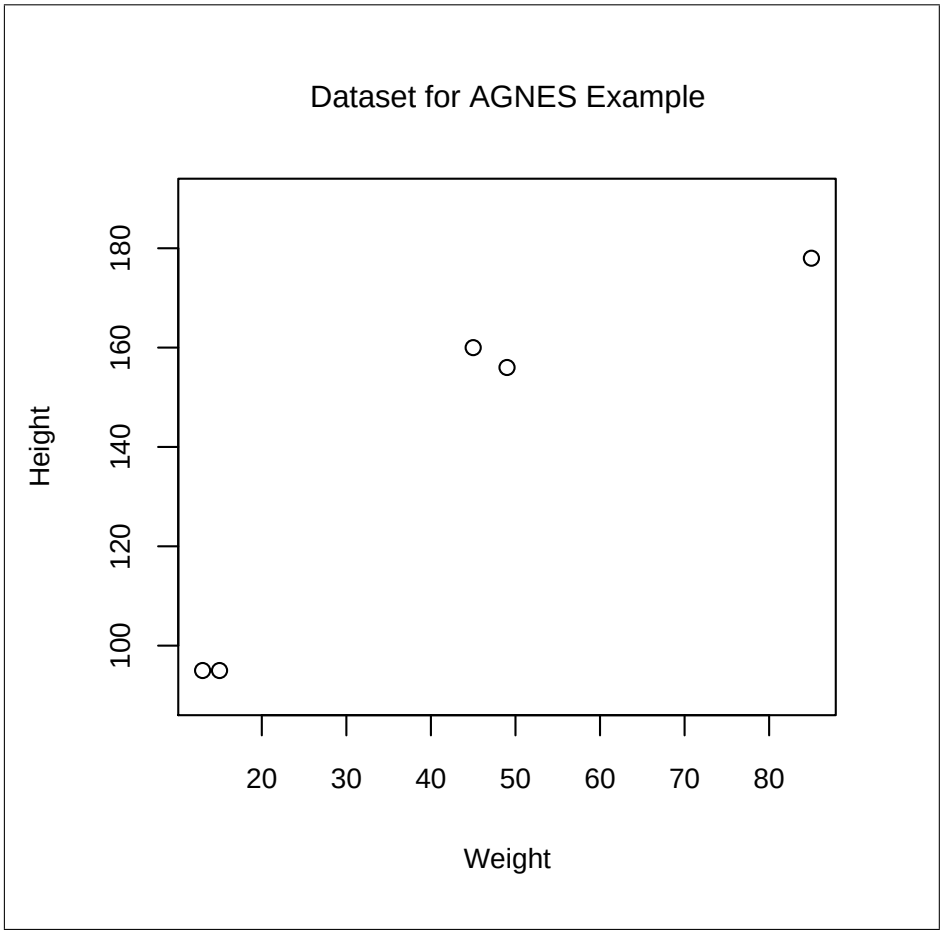

Figure 2–1: AGNES Example Dataset Plot

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 2–3).

Looking at the Dissimilarity Matrix (Table 2–3) the nearest objects are Bryan and Joshua. Then, we join it and now we have four cluster: (1) {*Joshua*, *Bryan*}, (2) *Anne*, (3) *Peter*, and (4) *Cristin*.

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 2.4 and will obtain a new matrix. The new distances are:

Name	*Weight* (kg)	*Height* (cm)
Joshua	−0.8957	−1.0705
Anne	0.2578	0.4917
Bryan	−0.9635	−1.0705
Peter	0.1221	0.5941
Cristin	1.4792	1.0551

Table 2–2: Standardized Data, AGNES Example

	Joshua	Anne	Bryan	Peter	Cristin
Joshua	0.0000	1.9419	0.0679	1.9511	3.1872
Anne	1.9419	0.0000	1.9830	0.1700	1.3450
Bryan	0.0679	1.9830	0.0000	1.9874	3.2381
Peter	1.9511	0.1700	1.9874	0.0000	1.4332
Cristin	3.1872	1.3450	3.2381	1.4332	0.0000

Table 2–3: Proximity Matrix, AGNES Example

$$d(\{Joshua, Bryan\}, Anne) = \frac{1}{2}[d(Joshua, Anne) + d(Bryan, Anne)]$$
$$= \frac{1}{2}[1.94 + 1.98]$$
$$= 1.96$$

$$d(\{Joshua, Bryan\}, Peter) = \frac{1}{2}[d(Joshua, Peter) + d(Bryan, Peter)]$$
$$= \frac{1}{2}[1.95 + 1.99]$$
$$= 1.97$$

$$d(\{Joshua, Bryan\}, Cristin) = \frac{1}{2}[d(Joshua, Cristin) + d(Bryan, Cristin)]$$
$$= \frac{1}{2}[3.19 + 3.24]$$
$$= 3.215$$

The new proximity matrix is

	{Joshua, Bryan}	Anne	Peter	Cristin
{Joshua, Bryan}	0.00	1.96	1.97	3.215
Anne	1.96	0.00	0.17	1.35
Peter	1.97	0.17	0.00	1.43
Cristin	3.215	1.35	1.43	0.00

Now, using the new proximity matrix, we need to find the nearest clusters. We can see that the nearest ones are *Anne* and *Peter*, Then, the new clusters are: (1) {*Joshua*, *Bryan*}, (2) {*Anne*, *Peter*}, and (3) *Cristin*.

Step 3. We need to repeat the previous step with the new clusters. We will calculate the distance between the clusters and obtain a new proximity matrix.

The new distance are,

$$d(\{Joshua, Bryan\}, \{Anne, Peter\}) = \frac{1}{4}[1.94 + 1.95 + 1.98 + 1.99]$$
$$= 1.965$$

$$d(\{Anne, Peter\}, Cristin) = \frac{1}{2}[d(Anne, Cristin) + d(Peter, Cristin)]$$
$$= \frac{1}{2}[1.35 + 1.43]$$
$$= 1.39$$

and the new proximity matrix is,

	{Joshua, Bryan}	{Anne, Peter}	{Cristin}
{Joshua, Bryan}	0.00	1.965	3.215
{Anne, Peter}	1.965	0.00	1.39
{Cristin}	3.215	1.39	0.00

As in the previous step, we need to join the nearest clusters, that is, those who have the smaller distance. In this step {Cristin} and {Anne, Peter} are joined. Then the new clusters are: (1) {*Joshua*, *Bryan*}, (2){*Anne*, *Peter Cristin*}.

Step 4. This is the last step and the only thing to do is to join the two cluster. Then, in the last step all the vectors lies in a single cluster.

Because the AGNES algorithm is a hierarchical one, when we want to obtain a graph of the results, the only way to obtain it is doing a Dendrogram. For a dendrogram of our previous example see Figure 2–2.

2.3.2 Hierarchical Divisive Clustering (DIANA)

This methods work reciprocally of the one discussed on Section 2.3.1. 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 [10] [7]. The algorithm can be described as follows.

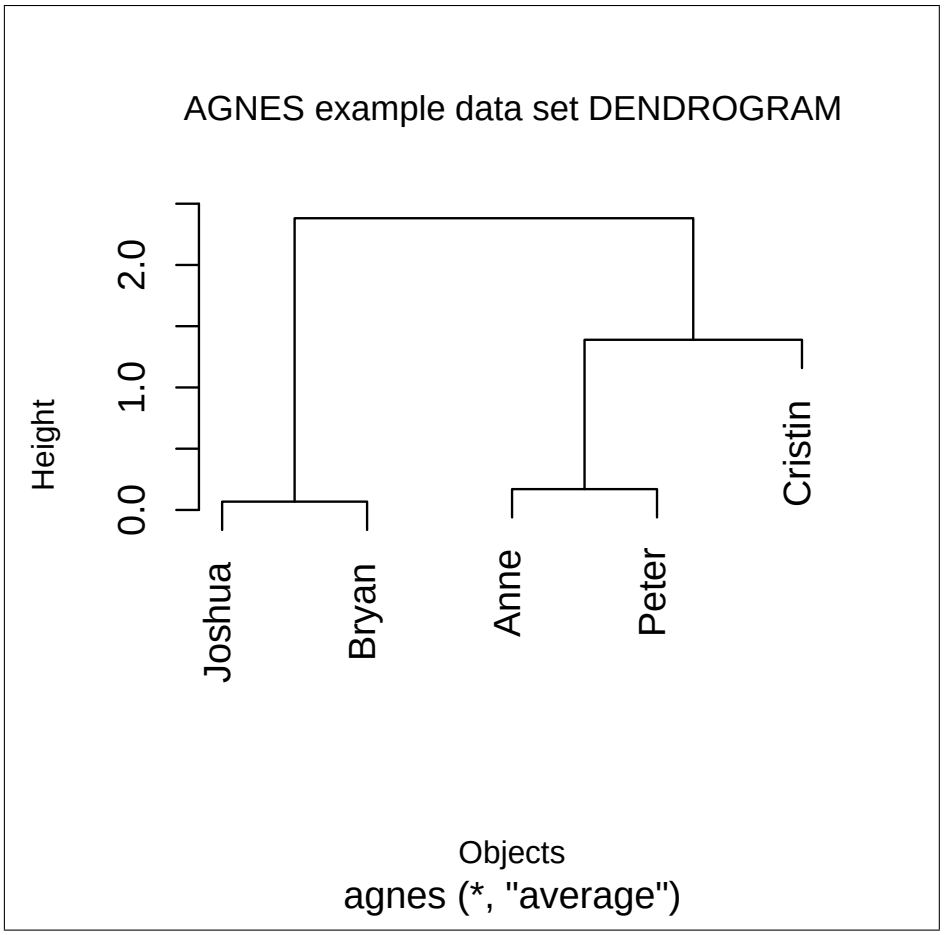

Figure 2–2: Dendrogram for AGNES Example

(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.

Now, to illustrate the Divisive Clustering Algorithms, lets do an example. Suppose we have the data set shown in Table 2–4.

Object	*Variable* 1	*Variable* 2
A	2	2
B	5.5	4
C	5	5
D	1.5	2.5
E	1	1
F	7	5
G	5.75	6.5

Table 2–4: Data set, DIANA Example

The first thing to do, as in the example in the previous section, is to standardize the data set. The standardized data for the DIANA example is shown in Table 2–5.

	Variable 1	Variable 2
A	−0.82	−0.88
B	0.64	0.15
C	0.43	0.66
D	−1.03	−0.62
E	−1.24	−1.39
F	1.26	0.66
G	0.74	1.43

Table 2–5: Standardized Data, DIANA Example

Once the data set have been standardized, the next step before do the clustering algorithms is to obtain the Proximity Matrix using the Euclidean Distance. Table 2–6 have the proximity matrix for this example.

Step 1. In the first step, the algorithm has to split the data set into two clusters. First we need to find the average dissimilarity from all the pairs of objects in order to find the most dissimilar to the others objects. Once we find the most

dissimilar, this object in particular, must be the one chosen to begin the splinter group. The splinter group is the one formed by the objects more dissimilar to the rest of the objects in the data set X. Table 2–7 contains the average dissimilarity measures from one object to the rest on the objects in X.

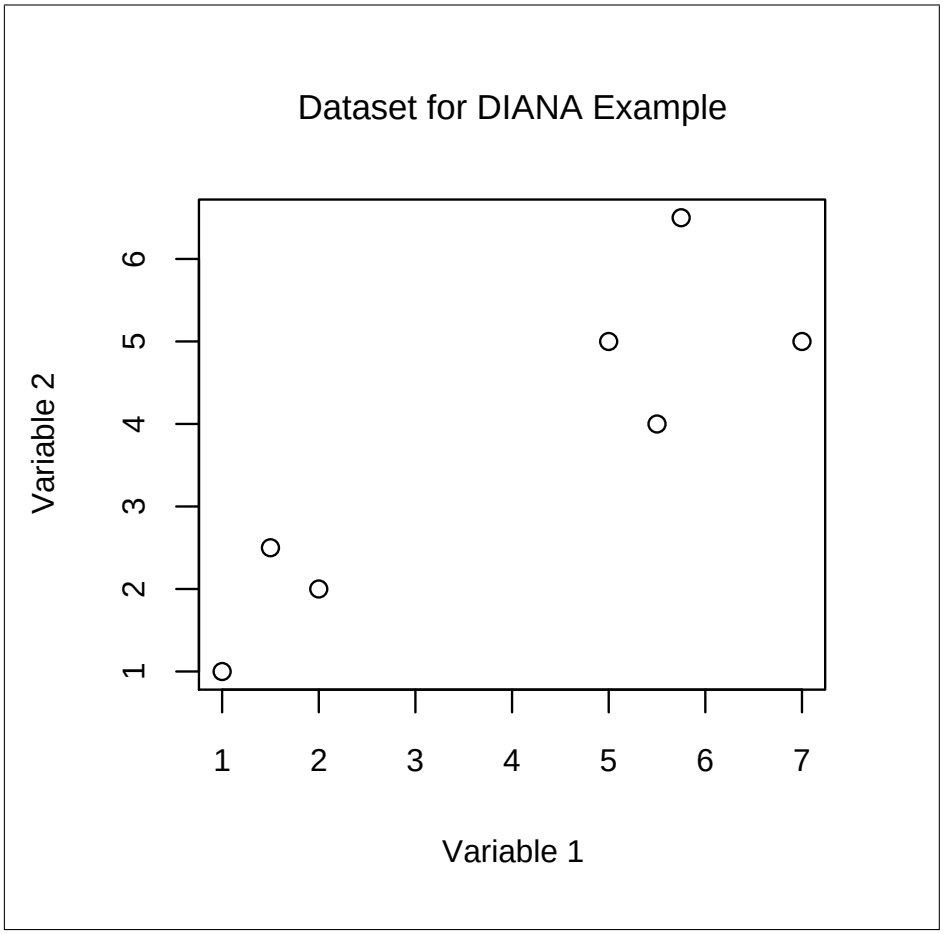

Figure 2–3: DIANA Example Data set Plot

	A	B	C	D	E	F	G
A	0.00	1.78	1.98	0.33	0.66	2.59	2.78
B	1.78	0.00	0.55	1.83	2.42	0.81	1.28
C	1.98	0.55	0.00	1.94	2.64	0.83	0.83
D	0.33	1.83	1.94	0.00	0.80	2.62	2.71
E	0.66	2.42	2.64	0.80	0.00	3.23	3.44
F	2.59	0.81	0.83	2.62	3.23	0.00	0.93
G	2.78	1.28	0.83	2.71	3.44	0.93	0.00

Table 2–6: Proximity Matrix, DIANA Example

Object	Average Dissimilarity to the Other Objects
A	$(1.78 + 1.98 + 0.33 + 0.66 + 2.59 + 2.78)/6 = 1.687$
B	$(1.78 + 0.55 + 1.83 + 2.42 + 0.81 + 1.28)/6 = 1.445$
C	$(1.98 + 0.55 + 1.94 + 2.64 + 0.83 + 0.83)/6 = 1.462$
D	$(0.33 + 1.83 + 1.94 + 0.80 + 2.62 + 2.71)/6 = 1.705$
E	$(0.66 + 2.42 + 2.64 + 0.80 + 3.23 + 3.44)/6 = 2.198$
F	$(2.59 + 0.81 + 0.83 + 2.62 + 3.23 + 0.93)/6 = 1.835$
G	$(2.78 + 1.28 + 0.83 + 2.71 + 3.44 + 0.93)/6 = 1.995$

Table 2–7: Dissimilarity Average

Observing Table 2–7, the most dissimilar object is E. So object E is chosen to initiate the splinter group. At this moment of the algorithm we have the groups: $\{E\}$ and $\{A, B, C, D, F, G\}$. Now we need to calculate the average dissimilarity with the remaining objects in the larger group, and compare it to the average dissimilarity with the object of the splinter group.

Object	Average Dissimilarity to Remaining Objects	Dissimilarity to Object of Splinter Group	Difference
A	$(1.78 + 1.98 + 0.33 + 2.59 + 2.78)/5 = 1.892$	0.66	1.232
B	$(1.78 + 0.55 + 1.83 + 0.81 + 1.28)/5 = 1.25$	2.42	-0.17
C	$(1.98 + 0.55 + 1.94 + 0.83 + 0.83)/5 = 1.226$	2.64	-1.414
D	$(0.33 + 1.83 + 1.94 + 2.62 + 2.71)/5 = 1.886$	0.80	1.086
F	$(2.59 + 0.81 + 0.83 + 2.62 + 0.93)/5 = 1.556$	3.23	-1.674
G	$(2.78 + 1.28 + 0.83 + 2.71 + 0.93)/5 = 1.706$	3.44	-1.734

The difference is largest for object A. Therefore, object A changes sides, so the new splinter group is {A, E} and the remaining group becomes {B, C, D, F, G}. Repeating the computations with the largest group we obtain:

Object	Average Dissimilarity to Remaining Objects	Average Dissimilarity to Objects of Splinter Group	Difference
B	$(0.55 + 1.83 + 0.81 + 1.28)/4 = 1.1175$	$(2.42 + 1.78)/2 = 2.1$	-0.9825
C	$(0.55 + 1.94 + 0.83 + 0.83)/4 = 1.0375$	$(2.64 + 1.98)/2 = 2.31$	-1.2725
D	$(1.83 + 1.94 + 2.62 + 2.71)/4 = 2.275$	$(0.80 + 0.33)/2 = 0.565$	1.71
F	$(0.81 + 0.83 + 2.62 + 0.93)/4 = 1.2975$	$(3.23 + 2.59)/2 = 2.91$	-1.6125
G	$(1.28 + 0.83 + 2.71 + 0.93)/4 = 1.4375$	$(3.44 + 2.78)/2 = 3.11$	-1.6725

Observing the differences, the largest one is obtained for object D. Therefore object D is moved to the splinter group. At this moment the new splinter group is {A, D, E} and the remaining group becomes {B, C, F, G}. When we repeat the computations we find:

Object	Average Dissimilarity to Remaining Objects	Average Dissimilarity to Objects of Splinter Group	Difference
B	$(0.55 + 0.81 + 1.28)/3 = 0.88$	$(2.42 + 1.78 + 1.83)/3 = 2.01$	-1.13
C	$(0.55 + 0.83 + 0.83)/3 = 0.7367$	$(2.64 + 1.98 + 1.94)/3 = 2.1867$	-1.45
F	$(0.81 + 0.83 + 0.93)/3 = 0.8567$	$(3.23 + 2.59 + 2.62)/3 = 2.8133$	-1.9566
G	$(1.28 + 0.83 + 0.93)/3 = 1.0133$	$(3.44 + 2.78 + 2.71)/3 = 2.9767$	-1.9634

After the calculations in the previous table, all the differences are negative, then the first step is concluded. In this way, the first step produced two clusters. The clusters are $\{A,\ D,\ E\}$, and $\{B,\ C,\ F,\ G\}$. Now, we need to find the diameters of the two clusters in order to select the cluster that have to be separated. The diameter of a cluster is the largest dissimilarity between the object belonging to the corresponding cluster. The diameter of the cluster $\{A,\ D,\ E\}$ is 0.80, and the

diameter of the cluster $\{B,\ C,\ F,\ G\}$ is 1.28. In the next step the cluster $\{B,\ C,\ F,\ G\}$ will be separated.

Step 2. To begin with this step, let see the dissimilarity matrix obtained using the objects in the cluster to be separated.

	B	C	F	G
B	0.00	0.55	0.81	1.28
C	0.55	0.00	0.83	0.83
F	0.81	0.83	0.00	0.93
G	1.28	0.83	0.93	0.00

In the sequel we need to calculate the average dissimilarity between the object in the cluster.

Object	Average Dissimilarity to the Other Objects
B	$(0.55 + 0.81 + 1.28)/3 = 0.88$
C	$(0.55 + 0.83 + 0.83)/3 = 0.7367$
F	$(0.81 + 0.83 + 0.93)/3 = 0.8567$
G	$(1.28 + 0.83 + 0.93)/3 = 1.0133$

As object $\{G\}$ has have the largest average dissimilarity to the other objects, therefore, it is the one chosen to begin the new splinter group. At this moment we have the groups $\{G\ \}$ and $\{B,\ C,\ F\}$. Now we need to calculate the average dissimilarity with the remaining objects in the larger group, and compare it to the average dissimilarity with the objects of the splinter group.

Object	Average Dissimilarity to Remaining Objects	Average Dissimilarity to Objects of Splinter Group	Difference
B	$(0.55 + 0.81)/2 = 0.68$	1.28	-0.6
C	$(0.55 + 0.83)/2 = 0.69$	0.83	-0.14
F	$(0.81 + 0.83)/2 = 0.82$	0.93	-0.11

As in the previous step all the differences are negatives, then the second step stop here. Therefore we have three clusters, {A, D, E}, {B, C, F}, and {G} (also called a singleton, with diameter 0).

In order to continue with the next step, we must decide which of these clusters to split. Let calculate the diameter of the clusters we have, and choose the larger one. The singleton {G} cannot be divided any further, and its diameter is 0. The cluster {A, D, E} has diameter 0.80, and the cluster {B, C, F} has diameter 0.83. Therefore, in the next step, we have to divide the cluster {B, C, F}.

Step 3 As in the previous step, we first observe at the dissimilarity matrix.

	B	C	F
B	0.00	0.55	0.81
C	0.55	0.00	0.83
F	0.81	0.83	0.00

Proceeding as in the previous steps, we obtain:

Object	Average Dissimilarity to the Other Objects
B	$(0.55 + 0.81)/2 = 0.68$
C	$(0.55 + 0.83)/2 = 0.69$
F	$(0.81 + 0.83)/2 = 0.82$

Then object {F} is the one chosen to begin the splinter group. Afterward, we obtain

Object	Average Dissimilarity to Remaining Objects	Average Dissimilarity to Objects of Splinter Group	Difference
B	0.55	0.81	−0.26
C	0.55	0.83	−0.28

The process stop because all differences are negative. Then, this step divides {B, C, F} into {F} and {B, C}. So, at the end of this step we have four clusters, {A, D, E}, {G}, {F}, and {B, C}.

To continue with the next step we must find a cluster to split. Lets calculate the diameter of these clusters. The cluster {A, D, E} has diameter 0.80, that of {B, C} is 0.55 and the other two have diameter 0. Therefore, the next cluster to be separated is {A, D, E}.

Step 4 The dissimilarity matrix for the objects in {A, D, E} is:

	A	D	E
A	0.00	0.33	0.66
D	0.33	0.00	0.80
E	0.66	0.80	0.00

Now, we need to find the most dissimilar one to be chosen to begin with the new splinter group.

Object	Average Dissimilarity to the Other Objects
A	$(0.33 + 0.66)/2 = 0.495$
D	$(0.33 + 0.80)/2 = 0.565$
E	$(0.66 + 0.80)/2 = 0.73$

Therefore object E is the one chosen to begin the new splinter group. Afterward, we find

Object	Average Dissimilarity to Remaining Objects	Average Dissimilarity to Objects of Splinter Group	Difference
A	0.33	0.66	-0.33
D	0.33	0.80	-0.47

This step conclude because all differences are negative. Then, this step divides {A, D, E} into {E} and {A, D}. So, at the end of this step we have five clusters, {A, D}, {E}, {G}, {F}, and {B, C}.

Now, calculate the diameters in order to find the next cluster to split. The diameters are 0.33, 0, 0, 0, and 0.55 respectively. So, the next cluster to be separate is {B, C}.

Step 5 The dissimilarity matrix for this cluster is:

	B	C
B	0.00	0.55
C	0.55	0.00

To find the one to start the splinter group, we compute

Object	Average Dissimilarity to the Other Objects
B	0.55
C	0.55

Because the average dissimilarities are the same, we may choose either object to begin the splinter group with [1] . Let us choose object B, and we obtain {B} and {C}. Therefore, we have two singletons that cannot be divided. So, this step divides {B, C} into two singletons {B} and {C}. At the end of this step we have six clusters, {A, D}, {E}, {G}, {F}, {B}, and {C}.

Step 6 In this step we have to split up the cluster {A, D}, because all the others are singletons. Proceeding as the previous step, lets see the dissimilarity matrix.

[1] When we have an object with the same largest dissimilarity measure we have to choose one at random [7]

	A	D
A	0.00	0.33
D	0.33	0.00

To find the one to start the splinter group, we compute

Object	Average Dissimilarity to the Other Objects
A	0.33
D	0.33

As in the previous step this cluster must be divide into two singletons {A} and {D}. At the end of this step we have seven clusters, {A}, {D}, {E}, {G}, {F}, {B}, and {C}.

After this step we have seven singletons cluster, so the algorithm stops, and we have each object in a single cluster.

As in the Agglomerative Nesting, the Divisive Analysis is a hierarchical clustering algorithm. Then, if we want to obtain a graph of the results we need to do a Dendrogram. A dendrogram of our previous example is shown in Figure 2–4.

2.4 Partitioning Clustering Algorithms

The nonhierarchical clustering methods are refer as Partitioning Clustering Algorithm [8] [9]. A partitional clustering obtain a single partition of the data set instead of a clustering structure, such as the dendrogram discussed in the previous section. The algorithm used in the Partitioning Clustering is based on the search of k representative objects among the objects of the data set. This k representative objects should represent various aspect of the structure of the data, and are often

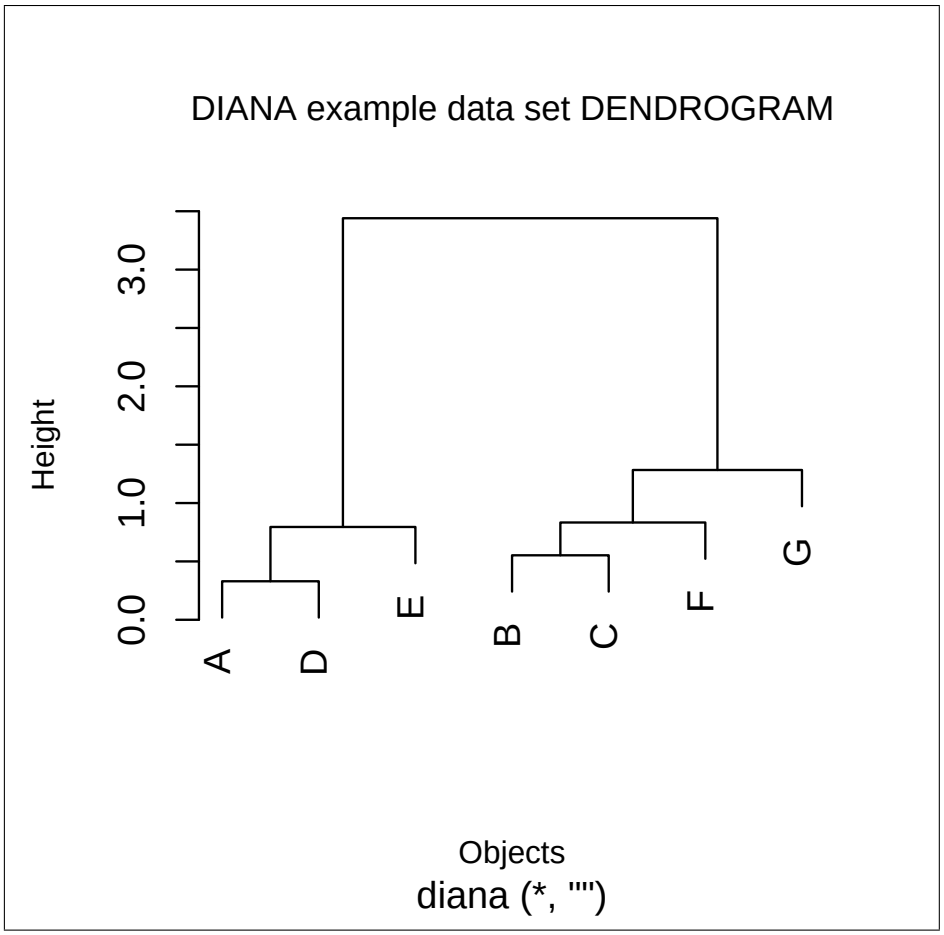

Figure 2–4: Dendrogram for DIANA Example

called *centrotypes*. After finding a set of k representative objects, the k clusters are constructed by assigning each object of the data set to the nearest representative object. Partitional methods have advantages when its applied to large data sets for which the construction of a Dendrogram is difficult. But, have disadvantages in the choice of the numbers of desired partitions. Two of these algorithms are discussed in this section: K - means, and Partitioning Around Medoids.

2.4.1 K-Means Clustering

K-means is a popular nonhierarchical clustering technique. In this case the k representative objects are called centroids. K-means is an iterative algorithms, and its basic idea is to start with an initial partition and assign observation to clusters so that the squared error decrease. The algorithm follows a simple way to classify a

given data set into k clusters fixed a priori. An iterative method can be implemented in different ways. Different implementation can lead to different partitions [11], [12] [9].

Typical convergence criteria are: no (or minimal) reassignment of patterns to new cluster centers, or minimal decrease in squared error. The algorithm proceeds as follows:

1. Select an initial k clusters centroid.
2. Assign each observation to its closest cluster centroid. That generate a new partition.
3. Compute the centroid of the new partition.
4. Repeat steps 2, and 3 until convergence is obtained. Typical convergence criteria are: no reassignment of patterns to new cluster centers, or minimal decrease in squared error.

There are several ways to choose the initial k centroid:

- Using the first k objects.
- Choosing k objects at random.
- Taking any partition of k cluster at random, and calculate its centroids.

Some important characteristic of the k-means algorithm are: (1) it is computationally fast, (2) it is sensible to outliers, and (3) can be performed with missing values. An example of this algorithm can be found in [8]. K-Means Clustering is not used in this research. It is just explained because of its relation with PAM, that will be discussed in next section. PAM is a generalization of the K-Means Clustering Algorithm.

2.4.2 Partitioning Around Medoids (PAM)

As we mention previously, this type of algorithms needs to find k representative objects. These k representative objects, in Partitioning Around Medoids (PAM)

algorithm, are called *medoids*. The best partition will be the one minimizing the average dissimilarity of objects to their closest representative object [7] [13].

The algorithms 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 set of representative objects that was chosen in the first phase. The algorithm works as follows:

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 its dissimilarity D_j with the first object chosen and calculate its dissimilarity $d(j,i)$ with object i. Calculate the difference between D_j and $d(j,i)$. If the difference is positive, then object j will contribute in the selection of object i. Then calculate,

$$C_{ji} = max(D_j - d(j,i), 0)$$

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

$$\sum_j C_{ji}$$

4. Choose the object i that maximizes

$$\sum_j C_{ji}$$

The process ends when the k representative objects have been found. Now, consider all the pair on objects (i,h) for which object i has 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 object i and h the following calculations need to be completed.

1. First, consider a object j that has non 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 ($d(j,i) = D_j$)
      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 than 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. Second, Add the contributions C_{jih} to calculate the total result of the swap:

$$T_{ih} = \sum_j C_{jih}$$

3. The next step will be to select the pair (i, h) which

$$minimizes_{i,h} T_{ih}$$

The swap is carried out if $minimum T_{ih}$ is negative and the algorithm return to step 1. If $minimum T_{ih}$ is positive or zero, then the swap is not carry out.

To illustrate the algorithm let us do an example[2] . Consider the data set in Table 2–8.

Object	x Coordinate	y Coordinate
A	1.00	4.00
B	5.00	1.00
C	5.00	2.00
D	5.00	4.00
E	10.00	4.00
F	25.00	4.00
G	25.00	6.00
H	25.00	7.00
I	25.00	8.00
J	29.00	7.00

Table 2–8: Data Set, PAM Example

First, suppose that the data set must be divided into two clusters (k=2). Now, we need to find the two medoids. First, we need to calculate the dissimilarity matrix of the data set (Table 2–9).

Next we need to find the object that has the smallest sum of dissimilarities to all the other objects. Table 2–10 shows the sum of dissimilarities for all the objects.

[2] This example was taken from [7]

	A	B	C	D	E	F	G	H	I	J
A	0.00	5.00	4.47	4.00	9.00	24.00	24.08	24.19	24.33	28.16
B	5.00	0.00	1.00	3.00	5.83	20.22	20.62	20.88	21.19	24.74
C	4.47	1.00	0.00	2.00	5.39	20.10	20.40	20.62	20.88	24.52
D	4.00	3.00	2.00	0.00	5.00	20.00	20.10	20.22	20.40	24.19
E	9.00	5.83	5.39	5.00	0.00	15.00	15.13	15.30	15.52	19.24
F	24.00	20.22	20.10	20.00	15.00	0.00	2.00	3.00	4.00	5.00
G	24.08	20.62	20.40	20.10	15.13	2.00	0.00	1.00	2.00	4.12
H	24.19	20.88	20.62	20.22	15.30	3.00	1.00	0.00	1.00	4.00
I	24.33	21.19	20.88	20.40	15.52	4.00	2.00	1.00	0.00	4.12
J	28.16	24.74	24.52	24.19	19.24	5.00	4.12	4.00	4.12	0.00

Table 2–9: Proximity Matrix, DIANA Example

	x
A	147.23
B	122.48
C	119.36
D	118.91
E	105.41
F	113.32
G	109.45
H	110.20
I	113.44
J	138.08

Table 2–10: Sum of dissimilarities

According to the previous table the object with the smallest dissimilarity, 105.41, is object E. Then object E is chosen as the first medoid. Now, in order to choose the second medoid, we need to follow the *Built Phase* of the algorithm.

Step 1 We need to consider an object i different to object E. This could be A, B, C, D, F, G, H, I, J. Let us consider $i = A$.

Step 2 Now, consider an object j that has not yet been selected. In this case object j could be B, C, D, F, G, H, I, and J. For each of them calculate its dissimilarity D_j with object E, and its dissimilarity $d(j, A)$ with object A. See Table 2–11.

j	$D_j = d(j, E)$	$d(j, A)$	$D_j - d(j, A)$	$C_{j,A} = max(D_j - d(j, A), 0)$
B	5.83	5.00	0.83	0.83
C	5.39	4.47	0.92	0.92
D	5.00	4.00	1.00	1.00
F	15.00	24.00	−9.0	0
G	15.13	24.08	−8.95	0
H	15.30	24.19	−8.89	0
I	15.52	24.33	−8.81	0
J	19.24	28.16	−8.92	0

Table 2–11: Contribution of object j to the selection of object A

Step 3 The total gain obtained by selecting object A is:

$$\sum_j C_{jA} = 2.75$$

The algorithm return to Step 1 and the calculations are made using objects B, C, D, F, G, H, I, and J. Now, we need to calculate the gain obtained by selecting each one of this objects. The results are shown below.

$$\sum_j C_{jB} = 10.39 \qquad \sum_j C_{jG} = 55.94$$

$$\sum_j C_{jC} = 12.36 \qquad \sum_j C_{jH} = 55.89$$

$$\sum_j C_{jD} = 11.22 \qquad \sum_j C_{jI} = 53.55$$

$$\sum_j C_{jF} = 51.19 \qquad \sum_j C_{jJ} = 43.71$$

Step 4 The second medoid is the one that $maximazes_i \sum_j C_{ji}$. Then, the second medoid is object G.

The Built Phase is completed with the selection of object E and object G as the two medoids. The next phase in the algorithm tries to improve the clustering algorithm's performance by swapping the two selected medoids with objects that have not yet been selected. First, we will swap object E with all the objects that

have not yet been selected, in order to determine the effect of the swap on the value of the clustering.

Following the *Swap Phase* we need to consider all the pairs of objects (i, h) where object i has been selected, but object h has not. First, consider $i = E$ and h any of the objects that has not been selected.

Step 1 Consider object h = object A, and object j, a nonselected object, and calculate its contribution C_{jiA} to the swap. First, calculate the dissimilarities between object j and object A, object E and the other medoid, object G.

j	$d(j,h)$ $d(j,A)$	$d(j,i)$ $d(j,E)$	$d(j,G)$ $d(j,G)$	C_{jih} C_{jEA}
B	5.00	5.83	20.62	−0.83
C	4.47	5.39	20.40	−0.92
D	4.00	5.00	20.10	−1.00
F	24.00	15.00	2.00	0
H	24.19	15.30	1.00	0
I	24.33	15.52	2.00	0
J	28.16	19.24	4.12	0

Table 2–12: Contribution of object j to the swap between objetcs

Objects F, H, I, and J are closer to the other medoid than from both object E and object A, then their contribution C_{jiA} is zero. Objects B, C, and D are closer to object A than to the other medoid, that is, $d(j, A) < d(j, G)$, then their contribution to the swap is calculated by $C_{jEA} = d(j, A) - d(j, E)$. The result can be observed in Table 2–12.

Step 2 The total result of the swap between Object E and object A is:

$$T_{EA} = \sum_j C_{jEA} = -2.75$$

Similarly,

$$T_{EB} = -10.39 \qquad T_{EH} = 30.23$$
$$T_{EC} = -12.36 \qquad T_{EI} = 30.57$$
$$T_{ED} = -11.22 \qquad T_{EJ} = 59.99$$
$$T_{EF} = 59.99$$

Step 3 In this step the pair (E, C) is selected because is the one that

$$minimizes_{i,h} T_{ih}.$$

In this case, the minimum has a negative value, then the swap between object E and object C is carried out. The algorithm return to Step 1 taking $i = G$ and finding T_{Gh} for all the pairs (G, h). After all the calculation are done, the results are the following,

$$T_{GA} = 24.22 \qquad T_{GH} = -0.12$$
$$T_{GB} = 19.76 \qquad T_{GI} = 29.42$$
$$T_{GD} = 18.24 \qquad T_{GJ} = 8.12$$
$$T_{GF} = 4.88$$

Again, in this case the minimum has a negative value, then the swap is carry out between object G and object H. The algorithm return to Step 1 taking $i = C$

and finding T_{Ch} for all (G, h) pairs. The results are shown next.

$$T_{CA} = 6.00 \qquad T_{CF} = 20.35$$

$$T_{CB} = 1.53 \qquad T_{CI} = 57.82$$

$$T_{CD} = 1.47 \qquad T_{CJ} = 57.82$$

In this case the minimum is obtained by swapping object C and object D, but it has a positive value. Then, the swap is not carried out and the algorithm is stopped. The algorithm ends with the selection of object C and object H as the two medoid of the data set.

CHAPTER 3
CLUSTER VALIDATION TECHNIQUES

3.1 Introduction

The majority of the clustering algorithms impose a clustering structure on the data set X, even though X may not posses such a structure. Because of that, we must have a measure that indicate us if the vectors of X form clusters before we apply a clustering algorithm. The problem of verifying whether X possesses a clustering structure is known as *Clustering Tendency*.

Once we assume that X possess 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.

Let C denote the clustering structure resulting after applying a clustering algorithm to X. In order to investigate cluster validity, there are three approaches [4]. The first is based in *external criteria*, the second in *internal criteria*, and the third in *relative criteria*. These are discuss in Sections 3.3, 3.4, and 3.5 respectively.

3.2 Hypothesis Testing in Cluster Validity

First, lets review *Hypothesis Testing*. Let H_0 and H_a be the null and the alternative hypotheses, respectively.

$$H_0 : \theta = \theta_0$$

$$H_a : \theta \neq \theta_0$$

Let R_ρ denote the critical region corresponding to significance level α of a test statistic q. Let ω be the set of all possible values that θ may take under H_a. The probability that q lies in a critical region, when $\theta \epsilon R_\rho$, is $W(\theta)$ and is defined as:

$$W(\theta) = P(q \epsilon R_\rho | \theta \epsilon \omega) \tag{3.1}$$

This function is called the power function and can be used for the comparison of two different statistical test. There are two types of errors that are associates with a statistical test.

- Type 1 Error - This error is made when H_0 is rejected when it is true. In this case the probability of a type 1 error is denoted by α, where α is called the significance level of the test. The probability of not rejecting H_0 when it is true is $1 - \alpha$.
- Type 2 Error - This error is made when H_0 is not rejected when it is false. The probability of a type 2 error is $1 - W(\theta)$.

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 [4]. The goal is now twofold:

- First, we most generate a reference data population under the hypothesis of random structure,
- Second, we must define an appropriate statistic, whose values are indicative of the structure of a data set, and compare the value that results from our data set X against the value obtained from the reference population.

There are different ways to generate the reference population: (1) Random position hypothesis, (2) Random graph hypothesis, and (3) Random label hypothesis. In this thesis we are only considering random position hypothesis.

- *Random position hypothesis.* This hypothesis is appropriate for ratio data. It can be used with either external or internal criteria. In the case of *Internal Criteria*,

the statistic q is defined as to measure the degree to which a clustering structure, produced by a clustering algorithm, and matches the proximity matrix of the corresponding data set. In *External Criteria* the statistic q is defined as to measure the degree of correspondence between a pre-specified structure imposed on X and the clustering that result after the application of a specific clustering algorithm to X. In both cases H_0 is rejected if the value of q is unusually small or large. The external and internal criteria are discussed in Section 3.3 and Section 3.4, respectively.

- *Random graph hypothesis.* Usually used when information that concern only the vectors themselves, or their relationship, are available. It is appropriate when ordinal proximities between vectors are used. Consider A, a $N \times N$ rank order symmetric matrix with zero diagonal elements (if dissimilarity measurements are used), and with integers in the range $[1, N(N-1)/2]$ as upper diagonal elements. The entry $A(i, j)$ represent the dissimilarity between objects i and j. For example, if $A(3, 7) = 5$ y $A(3, 1) = 10$ we can conclude that object 3 is more similar to object 7 than to object 1. Under the random graph hypothesis, the reference population consist of $N \times N$ rank order proximity matrices A_i with no ties, generated by inserting randomly the integers in the range $[1, N(N-1)/2]$ in its upper diagonal entries. The statistic q can be defined so as to measure the agreement between a rank order (proximity) matrix and the corresponding clustering structure. If the value of q is unusually small or large, then H_0 is rejected.
- *Random label hypothesis.* This hypothesis consider all the possible partitions of X into m groups and assumes that these all partitions are equally likely. The statistic q can be defined so as to measure the degree to which information inherent in the data set X matches a specific partition. Once more, H_0 is rejected if the value of q is unusually small or large.

3.3 External Criteria Measures

External criteria are used either (a) for a comparison of a clustering structure C, produced by a clustering algorithm, with a partition P of X drawn independently from C or (b) for measuring the degree of agreement between a predetermined partition P and the proximity matrix, PM, of X.

Comparison of P with a Clustering C

In this case we need 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 $C = \{C_1,...,C_m\}$ and $P = \{P_1,...,P_s\}$. 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 weather or not this pair of vectors belong to the same cluster or partition [4] [14] [13]. Let us define the following notation:

- **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.

Lets define $a = SS$, the number of pairs of vectors in X that belong to the same cluster in C and to the same group in P; $b = SD$, the number of pairs of vectors in X that belong to the same cluster in C and to different groups in P; $c = DS$, the number of pairs of vectors that belong to different clusters in C and to the same group in P; and $d = DD$ the number of pairs of vectors that belong to different clusters in C and to different groups in P. Then $a + b + c + d = M$, where M is the total number of possible pairs in X, $M = N(N-1)/2$. Let $m_1 = a + b$ be the number of pairs of vectors that belong to the same cluster in C, and $m_2 = a + c$ be the number of pairs of vectors that belong to the same group in P. Using these

definitions, we can define statistical indices that help us to measure how similar C and P are.

3.3.1 Rand Index

The Rand Index measures the fraction of the total number of pairs that are either in the same cluster and in the same partition, or in different clusters and in different partitions [4] [14]. The Rand index is defined as:

$$R = \frac{(a+d)}{M} \tag{3.2}$$

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

3.3.2 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 [4] [14]. The Jaccard Coefficient is defined as

$$J = \frac{a}{a+b+c} \tag{3.3}$$

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

3.3.3 Fowlkes-Mallows 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 [15] [14]. The FM Index is defined as:

$$FM = \sqrt{\frac{a}{a+b}\frac{a}{a+c}} = \frac{a}{\sqrt{m_1 m_2}} \tag{3.4}$$

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

3.3.4 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 [4] [14] [16], where $X(i,j)$ equal to 1 if the pair of vectors (x_i, x_j) belong to the same cluster in C and 0 otherwise, and $Y(i,j)$ equal to 1 if the pair of vector (x_i, x_j) belong to the same group in P and 0 otherwise. The statistic is defined as follows:

$$\Gamma = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} X(i,j)Y(i,j) \tag{3.5}$$

where $X(i,j)$ and $Y(i,j)$ are the (i,j) elements of the matrices X and Y, respectively. In order to obtain values between -1 and 1 we have the normalized version, denoted by $\hat{\Gamma}$ and defined as:

$$\hat{\Gamma} = \frac{\frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (X(i,j) - \mu_x)(Y(i,j) - \mu_y)}{\sigma_x \sigma_y} \tag{3.6}$$

where μ_x, μ_y, σ_x, and σ_y are the respective means and standard deviations. That is, $\mu_x = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (X(i,j))$, and $\sigma_x^2 = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (X(i,j))^2 - \mu_x^2$, and similarly we defined μ_y and σ_y. Equation 3.6 can be written as:

$$\hat{\Gamma} = \frac{\frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (X(i,j)Y(i,j)) - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{3.7}$$

Using the notation introduced in Section 3.3, and the fact that X and Y have a binomial distribution, we have that:

$$a + b = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (X(i,j)) \tag{3.8}$$

$$a + c = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (Y(i,j)) \tag{3.9}$$

$$a = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (X(i,j)Y(i,j)) \tag{3.10}$$

where a is the total number of pairs of vectors that belong to the same cluster in C and to the same group in P. Now, if we replace these identities we obtain that:

$$\left.\begin{aligned}
\mu_x &= \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (X(i,j)) = \frac{a+b}{M} \\
\mu_y &= \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (Y(i,j)) = \frac{a+c}{M} \\
\sigma_x^2 &= \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (X(i,j))^2 - \mu_x^2 = \frac{a+b}{M} - \left(\frac{a+b}{M}\right)^2 \\
\sigma_y^2 &= \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (Y(i,j))^2 - \mu_y^2 = \frac{a+c}{M} - \left(\frac{a+c}{M}\right)^2
\end{aligned}\right\} \tag{3.11}$$

Now, an equivalent formula to Hubert's $\hat{\Gamma}$ Statistic can be obtain by replacing Equation 3.11 in Equation 3.7, and obtain an equivalent [1] :

$$\hat{\Gamma} = \frac{Ma - (a+b)(a+c)}{\sqrt{(a+b)(a+c)(M-(a+b))(M-(a+c))}} \tag{3.12}$$

In order to show how these indices are calculate let us make an example. This example was done following the one in [4].

[1] This proof can be found in http://math.upr.clu.edu/ edgar/ESMA683505.htm

	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	z_{10}	z_{11}	z_{12}
z_1		DS	DD	DD	SS	DD	DD	DS	DD	DD	SS	DD
z_2			DD	DD	DS	DD	DD	SS	DD	DD	DS	DD
z_3				SS	DD	DD	DS	DD	SS	DD	DD	DS
z_4					DD	DD	DS	DD	SS	DD	DD	DS
z_5						DD	DD	DS	DD	DD	SS	DD
z_6							DD	DD	DD	SS	DD	DD
z_7								DD	DS	DD	DD	SS
z_8									DD	DD	DS	DD
z_9										DD	DD	DS
z_{10}											DD	DD
z_{11}												DD
z_{12}												

Table 3–1: External Criteria Example

Consider a data set, Z, consisting of twelve objects,

$$Z = \{z_1, z_2, z_3, z_4, z_5, z_6, z_7, z_8, z_9, z_{10}, z_{11}, z_{12}\}$$

Consider $C = \{\{z_1, z_5, z_{11}\}, \{z_2, z_8\}, \{z_6, z_{10}\}, \{z_3, z_4, z_9\}, \{z_7, z_{12}\}\}$ and $P = \{\{z_1, z_5, z_{11}, z_2, z_8\}, \{z_6, z_{10}\}, \{z_3, z_4, z_7, z_9, z_{12}\}\}$. The following table shows the type of all pairs of vectors in Z.

According to Table 3–1, $a = 9$, $b = 0$, $c = 12$, and $d = 45$. Now, we can use Equations 3.2, 3.3, 3.12, and 3.4 in order to calculate these indices.

- $Rand = \frac{a+d}{M} = \frac{9+45}{66} = 0.8182$
- $Jaccard = \frac{a}{a+b+c} = \frac{9}{9+0+12} = 0.4285$
- $Fowlkes\ and\ Mallows = \sqrt{\frac{a}{a+b}\frac{a}{a+c}} = \sqrt{\frac{9}{9+0}\frac{9}{9+12}} = 0.19639$
- $Hubert = \frac{Ma-(a+b)(a+c)}{\sqrt{(a+b)(a+c)(M-(a+b))(M-(a+c))}} = \frac{(66)(9)-(9+0)(9+12)}{\sqrt{(9+0)(9+12)(66-(9+0))(66-(9+12))}} = 0.5816$

3.4 Internal Criteria Measures

Using Internal Criteria, we are going to verify whether the clustering structure produced by a clustering algorithm fit the data, but using only information inherent to the data set.

3.4.1 The Davies-Bouldin Index

Let s_i be a 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 [17] [16] [13]:

- $R_{ij} \geq 0$
- $R_{ij} = R_{ji}$
- 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 an R_{ij} that satisfies these conditions is [7]:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{3.13}$$

Then the Davies-Bouldin Index is defined as

$$DB_m = \frac{1}{m} \sum_{i=1}^{m} R_i \tag{3.14}$$

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

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

$$d_{ij} = \|\bar{x}_i - \bar{x}_j\| = \sqrt{\sum_{k=1}^{l} |\bar{x}_{ik} - \bar{x}_{jk}|^2} \tag{3.15}$$

And the dispersion of a cluster C_i is defined as:

$$s_i = \sqrt{\frac{1}{n_i} \sum_{x \epsilon C_i} \|x - \bar{x}_i\|^2} \tag{3.16}$$

The DB_m is the average similarity between each cluster and its most similar one. Small values of DB are indicative of the presence of compact and well-separated clusters.

3.4.2 The Dunn Index

The Dunn Index is defined as [18] [16]:

$$D_m = \min_{i=1,\ldots,m} \{ \min_{j=i+1,\ldots,m} (\frac{d(C_i, C_j)}{\max_{k=1,\ldots,m} \operatorname{diam}(C_k)}) \} \tag{3.17}$$

where the dissimilarity function between two clusters C_i and C_j is

$$d(C_i, C_j) = \min_{x \epsilon C_i, y \epsilon C_j} d(x, y) \tag{3.18}$$

and the *diameter of a cluster* C is defined as

$$\operatorname{diam}(C) = \max_{x,y \epsilon C} d(x, y) \tag{3.19}$$

If X contains compact and well-separated clusters, Dunn's Index will be large, since the distance between clusters is expected to be large and the diameter of the cluster is expected to be small.

3.4.3 Silhouette Index

The Silhouette Index is useful when it is seeking compact and clearly separated clusters [19] [20]. In order to construct silhouettes we need a partition obtained by the application of some clustering algorithm, and the proximity matrix containing all the proximities between objects. For a given cluster, this method assign to each object of the cluster a quantitative measure $s(i)$, known as the silhouette width [21]. 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 object in cluster C_j. Consider any cluster C_k different to cluster C_j, and compute $b(i) = \min_{C_k \neq C_j} d(i, C_k)$ $(k = 1, 2, ..., c; k \neq j)$. Then, the silhouette width is defined as,

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3.20}$$

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 represent the second best choice for object i.

From the definition we can see that $-1 \leq s(i) \leq 1$. 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.

If the data consist of similarities and $a'(i)$ and $d'(i, C)$ represent the corresponding average similarities, then $b'(i) = max_{C \neq A} d'(i, C)$ [19] [3]. The interpretation is in the same way as before. Now, the silhouette width is defined as,

$$s(i) = \frac{a'(i) - b'(i)}{\max\{a'(i), b'(i)\}} \tag{3.21}$$

There is possible to calculate a cluster silhouette S_j, called *average silhouette width*, that represent the heterogeneity of cluster C_j [19] [3]. This quantitative measure can be obtained using:

$$S_j = \frac{1}{m} \sum_{i=1}^{m} s(i) \tag{3.22}$$

We can also consider a overall or *global silhouette width* denoted by GS_u, and define as:

$$GS_u = \frac{1}{c} \sum_{j=1}^{c} S_j \tag{3.23}$$

where U is any partition, $U \longleftrightarrow C : C_i \cup C_2 \cup ... \cup C_c$. This global silhouette value is used as a validity index for U. In order to choose the optimal number of clusters for a data set using this index, choose the partition U with the maximum GS_u.

When hierarchical methods are used to divide a data set in groups, there exists a graphical display, called dendrograms, that give a visual interpretation of the results obtained by the clustering algorithm. On the other hand, when partitioning technique is used the result of the clustering is a list of clusters with their objects, and it is not possible to construct a dendrogram to obtain some visual help in the interpretation. In 1987, Rousseeuw proposed a graphical display for the partitional methods in his paper *Silhouettes: a graphical aid to the interparatation and validation of cluster analysis* [19]. This graphical display is called a Silhouette Plot, and is used to obtain a visual interpretation of the result.

For example, refer to the data set in Table 2–8. In the statistical package R, we can obtain the silhouette width for each cluster, and the global silhouette width for a partition, using the function *silhouette*. In this case we divided the data in two clusters. Then we can make a plot of this silhouette width values, $s(i)$. Figure 3–1 illustrate the silhouette plot resulting using data set in Table 2–8.

From the silhouette plot we can conclude that both clusters are compact and clearly separated because the $s(i)$ values, 0.78 and 0.86 respectively, are close to 1.

3.5 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 clusterings, 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. So, the problem can be stated as: "Among the clusterings 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 [4] [16]:

- *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

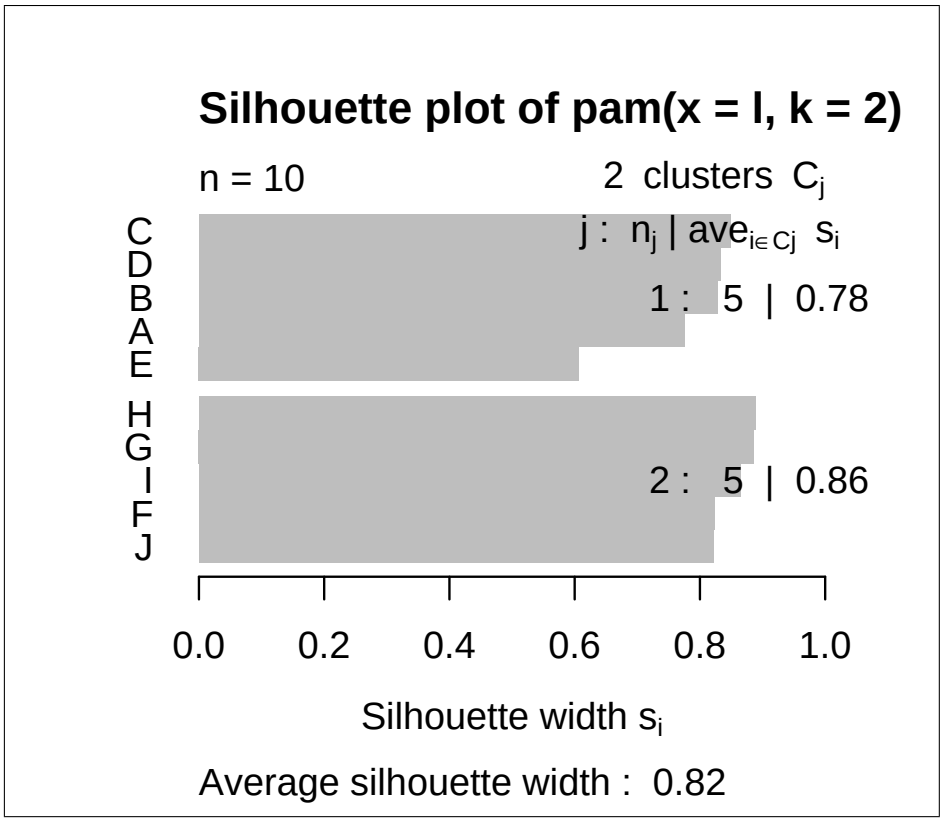

Figure 3–1: Silhouette Plot Example

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 nc_{max} and nc_{min}, 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 indicates good clustering.

Relative criteria measures are not considered in this thesis.

CHAPTER 4
EXPERIMENTAL RESULTS

In this work, three clustering algorithms have been used. Agglomerative Nesting Clustering algorithm was ran using all the possible combinations between metric and linkage method discussed in Chapter 3. There is a total of ten possible combinations, see Table 4–1. Divisive and Partitional Algorithm were ran using both metric, the Euclidean and Manhattan Distance. The validation indices used were: Davies-Bouldin Index, Dunn Index, Silhouette Index, Rand Statistic, Jaccard Coefficient, Fowlkes and Mallows Index, and Hubert Statistic.

Metric	Linkage Method	Notation
Euclidean (Eucl.)	Average	$IndexName_{11}$
	Single	$IndexName_{12}$
	Complete	$IndexName_{13}$
	Ward	$IndexName_{14}$
	Weighted	$IndexName_{15}$
Manhattan (Manh.)	Average	$IndexName_{21}$
	Single	$IndexName_{22}$
	Complete	$IndexName_{23}$
	Ward	$IndexName_{24}$
	Weighted	$IndexName_{25}$

Table 4–1: Combinations between Metric and Linkage Method

4.1 The Databases

A brief description of the data sets that were used in this thesis is given next, and a summary of its characteristics appears in Table 4–2.

1. ***Iris Plant*** (Iris): This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant: Iris Setosa, Iris Versicolour, and Iris Virginica. One class is linearly separable from the others two; the latter are NOT linearly separable from each other.
2. ***Wisconsin Breast Cancer*** (Breastw): This breast cancer databases was obtained from Dr. William H. Wolberg at the University of Wisconsin Hospital, Madison. Dr. Wolberg periodically reports his clinical cases from 1989 to 1991. It contains 699 instances, and each instance has one of 2 possible classes: benign or malignant. There are 16 instances containing missing values.
3. ***Johns Hopkins University Ionosphere database*** (Ionosphere): This radar data was collected by a system in Goose Bay, Labrador. This system consists of a phased array of 16 high-frequency antennas with a total transmitted power on the order of 6.4 kilowatts. The targets were free electrons in the ionosphere. "Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those that do not; their signals pass through the ionosphere. Returned signals were processed using an autocorrelation function whose arguments are the time of a pulse and the pulse number. There were 17 pulse numbers for the Goose Bay system. Instances in this database are described by two attributes per pulse number, corresponding to the complex values returned by the function resulting from the complex electromagnetic signal. The database contain 351 instance and 2 classes: "Good", and "Bad".
4. ***Breast cancer*** (Breastcc): This breast cancer data set contains the expression levels of 3227 genes for breast cancer patients with one of the three tumor types: sporadic, BRCA1 and BRCA2. It is described in Hedenfalk et al. (2001).

5. ***SRBCT***: This gene expression data set is presented in Kahn et al. (2001). It contains the expression levels of 2308 genes for 63 Small Round Blue Cells Tumor (SRBCT) patients belonging to one of the 4 tumor classes: Ewing family of tumors (EWS), non-Hodgkin lymphoma (BL), neuroblastoma (NB) and habdomyosarcoma (RMS).
6. ***Brain*** tumor: This dataset, presented by Pomeroy (2002), contains $5,597$ genes expression level of 42 microarray gene expression profiles from $K = 5$ different tumors of the central nervous system, that is, 10 medulloblastomas, 10 malignant gliomas, 10 atypical teratoid/rhabdoid tumors (AT/RTs), 8 primitive neuroectodermal tumors (PNETs) and 4 human cerebella.
7. ***Leukemia***: This data set is introduced by Golub et al. (1999) and contains the expression levels of 7129 genes for 47 ALL-leukemia patients and 25 AML-leukemia patients. It is included in the R library golubEsets. After data preprocessing following the procedure described in Dudoit et al. (2002), only 3571 variables remain. It is easy to achieve excellent classification accuracy on this data set, even with quite trivial methods as described in the original paper by Golub et al. (1999).
8. ***Lymphoma***: The data set presented by Alizadeh et al. (2000) comprises the expression levels of 4026 genes for 62 patients from 3 different classes: 42 samples diffuse large B-cell lymphoma (DLBCL), 9 samples from follicular lymphoma (FL), and 11 samples from chronic lymphocyte leukemia (CLL). The original data contains 96 samples and 9 classes.
9. ***Prostate***: This data set gives the expression levels of $12,600$ genes for 50 normal tissues and 52 prostate cancer tissues. We threshold the data and filter genes as described in Singh et al. (2002). The filtering step leaves us with 6033 genes.
10. ***Colon***: The colon data set is a publicly available 'benchmark' gene expression data set which is extensively described in Alon et al. (1999). The data set contains the expression levels of 2000 genes for 62 patients from two classes. 22 patients are

healthy patients and 40 patients have colon cancer. This data set is not as "easy" as the leukemia data set. The classification accuracy is usually much lower.

The first three data sets were obtained from the UCI Machine Learning Database Repository. The remaining data sets were obtained from the Dettling's website at http://stat.ethz.ch/ dettling/bagboost.html.

4.2 Internal Criteria Results

In this section the results from the internal indices are discussed. Davies-Bouldin Index results are presented in Section 4.2.1, Dunn Index results in Section 4.2.2, and the results using Silhouette Index are presented in Section 4.2.3

4.2.1 Davies - Bouldin Index

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 need to be large. According to this, Equation 3.14 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 cluster, then two of them will contain only one observation, and one cluster

Data Set	Number of Instances	Number of Classes	Instance per Classes	Number of Features
Iris	150	3	(50, 50, 50)	4
Breastw	683	2	(444,239)	9
Ionosphere	351	2	(225,126)	32
Breastcc	22	3	(7,8,7)	3226
SRBCT	63	4	(23,20,12,8)	2308
Brain	42	4	(10,10,10,4,8)	5597
Leukemia	72	2	(47,25)	3571
Lymphoma	62	3	(42,9,11)	4026
Prostate	102	2	(50,52)	6033
Colon	62	2	(22,40)	2000

Table 4–2: Data sets characteristics

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. Appendix B contains the tables with the results for DB Index.

Table B–1 shows the DB index obtained for Iris Data set. The minimum value among all the combinations using AGNES is obtained using Euclidean distance combined with SLINK method (DB_{12}) for $c = 3$ clusters. Using DIANA the minimum occur using Manhattan distance for $c = 2$, but for $c = 3$ there is a better clustering result. The same happened when using PAM. The clustering results are:

Cluster	AGNES	DIANA		PAM	
	$c=3$	$c=2$	$c=3$	$c=2$	$c=3$
1	49	50	50	50	50
2	1	100	42	100	57
3	100		58		43

Table 4–3: Clustering Results for Iris, using DB Index

Table B–2 shows the values obtained for Breastw Data set. Notice that, the minimum for each combination is obtained for $c = 2$ in most of the cases, indicating that Breastw must has two clusters. The minimum among all the combinations was obtained from DB_{21}. The patient classified benign breast cancer was totaly recognize, but the ones classified malign breast cancer are not. Again, the algorithm detect the two classes Breastw has, but they are not the same that the data set originally have. DIANA and PAM also recognized the two classes Breastw has. PAM seems to recognized better the two clusters. The clustering results are:

Cluster	AGNES	DIANA	PAM
1	652	507	461
2	183	176	222

Table 4–4: Clustering Results for Breastw, using DB Index

Table B–3 shows the values obtained for Ionosphere Data set. This data set has two classes. If we observe the table, the minimums are obtained for $c = 8$ using AGNES, $c = 6$ using DIANA with Euclidean distance, and $c = 2$ using PAM with Manhattan distance. In these cases the clustering result are:

Cluster	AGNES	DIANA	PAM
1	338	341	195
2	1	1	156
3	2	2	
4	2	1	
5	1	5	
6	3	1	
7	3		
8	1		

Table 4–5: Clustering Results for Ionosphere, using DB Index

Once again PAM gives good results.

Table B–4 shows the values obtained for Breastcc Data set. This data set contain 3227 genes classified in three types of cancer: Sporadic, BRCA1 and BRCA2. Observing the bold values in the table, there are no minimum values obtained for $c = 3$. This implies that the DB index always fails to detect the three clusters Breastcc have. In this case, DB index suggests that Breastcc could have two classes instead of three.

Cluster	AGNES	DIANA	PAM
1	20	20	16
2	2	2	6

Table 4–6: Clustering Results for Breastcc, using DB Index

Table B–5 shows the DB values obtained for SRBCT Data set. SRBCT data set has four classes. There are no minimum number for any combination for $c = 4$. The minimum is obtained for $c = 2$ using AGNES and PAM. Using DIANA, DB index indicate $c = 3$ as the optimal number of clusters for the data set. The clustering results are:

Cluster	AGNES	DIANA	PAM
1	61	61	29
2	1	1	34
3		2	

Table 4–7: Clustering Results for SRBCT, using DB Index

For $c = 4$ PAM yields the best clustering result: Cluster1 (17), Cluster2 (20), Cluster3 (6), and Cluster4 (20).

Table B–6 shows the values obtained for Brain Data Set. This data set is divided into five clusters. As we can observe in the table, for the AGNES Algorithms, DB Index suggest that Brain Data set must have two classes instead of five. DIANA and PAM algorithm also fail to detect the clusters Brain Data set has. The minimum is 0.0164 and it is obtained when using AGNES algorithm with the combinations DB_{11} and DB_{15}. Observing the clustering result for $c = 5$ using PAM, we obtain better results. The clustering result were:

Cluster	AGNES	DIANA	PAM
1	38	26	13
2	1	13	7
3	1	1	10
4	1	1	8
5	1	1	4

Table 4–8: Clustering Results for Brain, using DB Index

The values obtained using the Leukemia data set are shown in Table B–7. This data set consists of 3571 genes divided into 2 clusters. Using DB index the optimal number of clusters for this data set is two using AGNES, DIANA, and PAM. The best result is obtained when using PAM.

Cluster	AGNES	DIANA	PAM
1	71	71	27
2	1	1	45

Table 4–9: Clustering Results for Leukemia, using DB Index

Table B–8 shows the values obtained for the Lymphoma data Set. Using this index, it is obtained that Lymphoma must have two clusters instead of three. But if we look at the results of PAM for $c = 3$, we obtain Cluster1 (37), Cluster2 (14), and Cluster3 (11), which is very close to the true values.

Cluster	AGNES	DIANA	PAM
1	61	40	39
2	1	22	23

Table 4–10: Clustering Results for Lymphoma, using DB Index

DB values obtained from Prostate data set are shown in Table B–9. Using this index Prostate data set must have two clusters. The better clustering is obtained using PAM. The two clusters obtained from AGNES, DIANA and PAM are:

Cluster	AGNES	DIANA	PAM
1	101	100	42
2	1	1	60

Table 4–11: Clustering Results for Prostate, using DB Index

The clusters of size may indicate the presence of outliers.

The results for the Colon data set when the DB index is applied are shown in Table B–10. Using this index, it is detected that Colon data set must have two clusters. These two clusters are the following:

Cluster	AGNES	DIANA	PAM
1	61	58	29
2	1	4	33

Table 4–12: Clustering Results for Colon, using DB Index

Summarizing the results Davies-Bouldin Index fails to detect the number of cluster the data set has. Sometimes the number of optimal clusters suggested by this index match the number of clusters the data set originally has, but with different objects that compound each cluster. In general, using DB Index, the clustering algorithm that better classified the data sets was PAM with Manhattan distance, and the worst results were obtained using AGNES.

4.2.2 Dunn Index Results

The Dunn Index measures how compact and well-separated the clusters are, as in the case of Davies-Bouldin Index. As mentioned previously, to obtain compact and well-separated clusters, the dispersion measure for each cluster needs to be as small as possible, while the dissimilarity measure between clusters need to be large. According to this, Equation 4.2.2 would have large values if the clusters are compact and well-separated. Tables in Appendix C show the Dunn Index results.

Table C–1 shows the values obtained from Iris data set. Using AGNES the maximum is 0.3389, and it is obtained from D_{12}, D_{14}, D_{23}, D_{25} for $c = 2$. The clustering result are the following: for D_{12} and $D_{14} \Rightarrow$ Cluster1 = 49, Cluster2 = 101; for $D_{23} \Rightarrow$ Cluster1 = 71, Cluster2 = 79; and for $D_{25} \Rightarrow$ Cluster1 = 50, Cluster2 = 100. Using DIANA and PAM also is detected that Iris would have two classes. The clustering results are the same one obtained from D_{25}. Then, using this index it is suggested that the optimal number of clusters for Iris data set are two.

The values for Breastw data set are shown in Table C–2. The maximum value is obtained from D_{13} and D_{14} for $c = 2$ using AGNES. The clustering results are: for $D_{13} \Rightarrow$ Cluster1 = 500, Cluster2 = 183; and for $D_{14} \Rightarrow$ Cluster1 = 425, Cluster2 = 258. Using DIANA the maximum is obtained for $c = 9$ for which the index value is 0.1773, but observing the value for $c = 2$ it is not so far from 0.1773. The clustering result for $c = 2$ using DIANA with Manhattan distance is: Cluster1 = 507, Cluster2 = 176. When PAM is used, Dunn Index suggests that the optimal number of clusters for Breastw data set will be three. This clustering result is as follow: Cluster1 = 448, Cluster2 = 199, Cluster3 = 36.

Table C–3 show the values obtained from Ionosphere data set. The maximum value is obtained from D_{11}, D_{13} and D_{14} for $c = 2$ using AGNES. The clustering results are: for $D_{11} \Rightarrow$ Cluster1 = 350, Cluster2 = 1; for $D_{13} \Rightarrow$ Cluster1 = 264, Cluster2 = 87; and for $D_{14} \Rightarrow$ Cluster1 = 192, Cluster2 = 159. Using DIANA the maximum value is 0.4204 for $c = 2$. This clustering result is a poor one because it has 344 objects in Cluster1 and 7 objects in Cluster2. On the other hand, there are two clusters detected using PAM. The results are: Cluster1 = 191, and Cluster2 = 160. AGNES and PAM give better results for this data set.

The values from Breastcc data set are shown in Table C–4. The maximum value obtained using AGNES is 0.8275 for $c = 9$ clusters. Using DIANA the maximum

value is 0.8195 for $c = 10$, and using PAM the maximum value is 0.7389 for $c = 9$. This index fails in detect the number of clusters the data set has.

4.2.3 Silhouette Index

Tables from D–1 to D–10 show the silhouette results using AGNES, DIANA, and PAM dividing each data set in different number of clusters (2 to 10). The results were obtained using the Euclidean and Manhattan distance. The entries in the tables represent the GS_u (Global Silhouette Width for a partition U). The GS_u is the average of the silhouette width $S(i)$ obtained for each cluster in partition U. In order to select the optimal number of clusters for a data set, choose the partition U in where GS_u is a maximum.

Table D–1 shows the silhouette result for Iris Data set. The maximum value obtained, using AGNES is 0.3389. This value is obtained from Si_{21}, Si_{22}, Si_{24}, and Si_{25} for $c = 2$. The same value was obtained using DIANA with Euclidean and Manhattan Distance and using PAM with Euclidean distance. The clustering result is the same for all of them: Cluster1 = 50, Cluster2 = 100. In this case, Silhouette index fails in detecting the number of clusters Iris data set has.

Table D–2 shows the silhouette result for Breastw Data set. The maximum value obtained using AGNES is 0.5867 from Si_{23} for $c = 2$. Using DIANA the maximum value is obtained using Manhattan distance for $c = 2$, and using PAM the maximum is also obtained using Manhattan distance for $c = 2$. The clustering results are: for $Si_{23} \Rightarrow$ Cluster1 = 452, Cluster2 = 231; for DIANA $\Rightarrow$ Cluster1 = 507, Cluster2 = 176; for PAM $\Rightarrow$ Cluster1 = 448, Cluster2 = 235. The index detect the clusters the data set has.

The values obtained for Ionosphere data set are shown in Table D–3. The maximum values are obtained from Si_{22} for $c = 2$, DIANA with Manhattan distance for $c = 2$ and PAM with Manhattan distance for $c = 4$. For Si_{22} the clustering result is Cluster1 = 350, Cluster2 = 1; and for DIANA Cluster1 = 344, Cluster2 = 7. Both

clustering results are not a good clustering, because there are so many objects in one cluster. Thus, this index does not detect the clusters this data set has.

The values obtained for Breastcc data set are shown in Table D–4. Using AGNES the maximum is obtained from Si_{12} for $c = 2$. The clustering result in this case is: Cluster1 = 21, Cluster2 = 1. When using DIANA it is detected that the data set must have nine clusters. Finally, using PAM there is a maximum with $c = 2$. The result is Cluster1 = 16, Cluster2 = 6. When observing what happened if c=3 is selected as the optimal number of clusters if using PAM, the clustering result obtained is: Cluster1 = 6, Cluster2 = 11, Cluster3 = 5. This result is similar to the original classification.

Table D–5 shows the values obtained for Srbct data set. Using AGNES the maximum value is obtained for $c = 10$, using DIANA for $c = 9$, and using PAM for $c = 5$. The index value need to be close to one in order to be a good clustering. In these cases the values are small, close to zero, indicating a bad clustering result. The index fail to detect the number of clusters the data set has.

The values obtained for Brain data set are shown in Table D–6. Using AGNES the maximum value is obtained for $c = 10$, using DIANA for $c = 10$, and using PAM for $c = 2$. The index fail in detect the number of clusters the data set has.

The values obtained for Leukemia data set are shown in Table D–7. The maximum using AGNES is obtained from Si_{11}, Si_{12}, Si_{15}, Si_{21}, Si_{22}, and Si_{25}. The clustering result for all of them is 71 objects to Cluster1 and 1 object to Cluster2. Using DIANA occurs the same clustering result. Using PAM the maximum is obtained when used the Manhattan distance for $c = 2$. The clustering result in this case is: Cluster1 = 26, Cluster2 = 46. The index detects the clusters the data set has when using PAM.

The values obtained for Lymphoma data set are shown in Table D–8. Using AGNES the maximum value is obtained for $c = 2$, using DIANA for $c = 2$, and using

PAM for $c = 2$. The clustering results are: for $Si_{23} \Rightarrow$ Cluster1 = 41, Cluster2 = 21; for DIANA $\Rightarrow$ Cluster1 = 40, Cluster2 = 22; for PAM $\Rightarrow$ Cluster1 = 39, Cluster2 = 23. The index fails in detect the number of clusters the data set has.

The values obtained for Prostate data set are shown in Table D–9. 0.3704 is the maximum value obtained using AGNES. It is obtained from Si_{13}, Si_{15}, and Si_{21} for $c = 3$, and the clustering result is Cluster1 = 88, Cluster2 = 13, Cluster3 = 1 for all of them. Using DIANA and PAM the maximum is obtained for $c = 2$. The clustering results are: Cluster1 = 88, Cluster2 = 22 for DIANA, and Cluster1 = 42, Cluster2 = 60 for PAM, which is close to the true grouping.

The values obtained for Colon data set are shown in Table D–10. The maximum value obtained using AGNES is 0.1620 (Si_{12}), and the clustering result is Cluster1 = 60, Cluster2 = 1, Cluster3 = 1. Using DIANA the maximum is obtained for $c = 2$ using Euclidean Distance. The clustering result is Cluster1 = 58, Cluster2 = 4. Using PAM the maximum is obtained for $c = 10$. Hence the silhouette index does not perform well.

4.3 External Criteria Results

The external criteria used in this thesis were the Rand Statistic, Jaccard Coefficient, Fowlkes and Mallows Index, and Hubert Statistic [1] . As discuss in Chapter 3, the external criteria can be use by comparing the structure C produced by a clustering algorithm with a partition P drawn independently. In our case, the data sets has a column that contains the class to which each object belong. Then, we will use this column as the partition P necessary to calculate the external indices. The values for the external indices lies between 0 and 1 in the case of Rand Statistic, Jaccard Coefficient, and Fowlkes and Mallows Index. For these indices the closer

[1] These indices were programmed by Edgar Acuña, Ph.D., University of Puerto Rico, Mayagüez

the value to 1, the better the agreement between the partition P and the results of the clustering algorithm as discussed in Chapter 3. On the other hand, the Hubert Statistic values are between -1 and 1. Let us see what happened when these indices are applied to the data sets Iris, Breastw, Ionosphere, Breastcc, Srbct, Brain, Leukemia, Lymphoma, Prostate, and Colon.

Tables E–1 to E–10 shows the values for the external indices for each data set. Table E–1 that contains the external results for the Iris data set. Using AGNES the larger values for the four indices are obtained using Manhattan distance with Ward linkage method for $c = 3$. The clustering results are:

- Cluster1 = 50. These 50 objects are classified together in the original partition.
- Cluster2 = 68, of which 49 are classified together in the original partition.
- Cluster3 = 32, of which 31 are classified together in the original partition.

Using DIANA it is always suggested that Iris must have four classes instead of three. The values obtained when it is used PAM with Manhattan distance are the largest ones. In this case the clustering result are: Cluster1 = 50, Cluster2 = 57, and Cluster3 = 43. Iris data set has three clusters, that are detected using AGNES with the combination 24 and PAM with Manhattan distance.

Continuing with the discussing of the results refer to Table E–2. This table contains the results for Breastw data set. For this data set the largest values for the four indices using AGNES algorithm are obtained with the combination Manhattan distance and Ward linkage method for $c = 2$. The clustering result is Cluster1 = 436 (429 objects are classified together in the original partition) and Cluster2 = 247 (232 objects belong to the same cluster originally). Using DIANA with Euclidean distance it is also selected $c = 2$ as the optimal number of clusters for this data set. The clustering result is Cluster1 = 520 (443 objects are classified together in the original partition) and Cluster2 = 163 (162 objects are classified together in the original partition). Finally, using PAM the maximum value is obtained for

$c = 2$ using Euclidean distance. The four indices indicates that Breastw must have two clusters. The clustering result in this case is Cluster1 = 448 (435 objects are classified together in the original partition) and Cluster2 = 235 (226 objects are classified together in the original partition). Breastw data set has two clusters, and are detected using AGNES with combination 24, DIANA, and PAM both with Manhattan distance.

Table E–3 shows the external results for Ionosphere data set. Using AGNES the only external index that detect the two clusters the data set has is FM. Nevertheless, the clustering result for $FM_{11} = FM_{12} = FM_{22}$ are poor because Cluster1 has 350 objects while Cluster2 has 1 object. If it is used DIANA with Euclidean distance, FM is the only index that has a maximum for $c = 2$. As happened in AGNES, the clustering result is poor. Now, using PAM the four indices suggest $c = 3$ as the optimal number of clusters that Ionosphere data set must have. Observing the clustering result, Cluster1 has 191 objects (157 belong to the same partition), Cluster2 has 122 objects (92 belong to the same partition), and Cluster3 has 38 objects (all belong to the same partition). Although PAM does not detect that this data set has two clusters, the three clusters found form a good partition. In this case FM is the index that most detect the clusters of the data set.

The external results for Breastcc data set are shown in Table E–4. For this data set R_{23} suggest $c = 9$ as the optimal number of clusters, J_{21} suggest $c = 6$, FM_{12} suggests $c = 2$, and H_{21} suggest $c = 6$. Using DIANA and PAM there are no index that suggest $c = 3$ as the optimal number of clusters for Breastcc data set. Then, the external indices fail to detect the number of clusters this data set has.

Observing Table E–5 the values for the external indices when using AGNES, DIANA, and PAM are not indicating that the data set must have four clusters. Rand and Hubert Statistics suggest that $c = 10$ is the optimal number of cluster for this data set using AGNES and DIANA. When using PAM it suggest $c = 8$. Jaccard

and FM Index suggest $c = 2$ when using AGNES and DIANA, but suggest $c = 8$ when using PAM.

The results for Brain data set are shown in Table E–6. For this data set the external indices does not work. Using AGNES, it is detected that the data set must have seven clusters with the indices J, FM, and H, using combination 14. Using DIANA and PAM all the indices indicates that the data set must have ten clusters.

Now, let us discuss the results obtained for Leukemia data set. The results are shown in Table E–7. When AGNES is used with Euclidean distance and Ward method, Jaccard and Rand index indicate that the data set has two clusters, Cluster1 = 34 and Cluster2 = 38. For the Average, Single or Weighted methods Cluster1 = 71 and Cluster2 = 1. The same happened if DIANA is used with Euclidean distance. Using PAM the clustering result is Cluster1 = 27 and Cluster2 = 45 with Euclidean distance. Then, the four indices detect the two clusters the data set has using AGNES with combination 14 and using PAM with Euclidean distance.

Lymphoma data set has three clusters. Table E–8 shows the external results obtained for this data set. Using AGNES with Euclidean distance and Ward linkage method, as well as Manhattan distance with Ward linkage method, indicates that the data set must have three cluster containing 40, 11, 11 objects per cluster respectively. Using DIANA it is detected by all the external indices that the data set must have two clusters instead of three. Finally, using PAM it is detected the presence of three clusters. These three clusters have 37, 14, and 11 objects respectively. The external indices detect in this data set the optimal number of clusters this data set must have using AGNES with combination 14 and 24, and using PAM with both metrics, Euclidean and Manhattan distances.

Table E–9 shows the external indices values for Prostate data set. This data set has two clusters. Using AGNES the maximum values is obtained for $c = 2$. The clustering result is Cluster1 = 101 and Cluster2 = 1. The same result is obtained

using DIANA with Euclidean distance. Using PAM with Euclidean distance, indices Jaccard and FM detect two clusters for the data set. The clustering result is Cluster1 = 42 and Cluster2 = 60. Then, only Jaccard coefficient and FM index detect the two clusters when it is used PAM with Euclidean Distance.

For Colon data set, the results are shown in Table E–10. Using AGNES the maximum for Jaccard and FM are obtained with the combinations 12 and 22. The clustering result is poor because in this case Cluster1 has 61 objects while Cluster2 has only 1 object. DIANA give a poor clustering too, Cluste1 = 58, and Cluster2 = 4. Using PAM the four indices indicates that Colon data set must have two clusters. The clustering result for PAM using Euclidean distance is Cluster1 = 29, and Cluster2 = 33.

The external indices seems to identify in a better way the optimal number of clusters a data set must have. In most of the cases in this work, external indices identify the number of clusters a data set have.

CHAPTER 5
CONCLUSIONS

5.1 Conclusions

There exists another works in which clustering validation techniques are used to evaluated clustering results [22] [23] [24]. In such works the data sets used were generated by simulation according to some criteria. The clusters were generated in the same manner. That is an important difference with this work because the data sets used here are real data sets.

In this thesis the main objective was to compare some of the validation indices and his precision in detecting the optimal number of classes a data set would have. Our empirical results show that the value of each index could be affected by the presence of outliers in the data set. When using AGNES the clustering result is affected by the outliers when using Average, Single, and Weighted linkage methods. The best results were obtained when using Complete and Ward Linkage Methods. But in general AGNES 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 General, PAM seems to do a better clustering task.

APPENDICES

APPENDIX A
CODES OF R FUNCTIONS USED IN THIS THESIS

Dunn Index Function

```
dunn = function(data,estimated) {
**************************************************
This function calculate the DUNN Index. You must run
a clustering algorithm before run this programm.
**************************************************
data - the data set
estimated - vector that contain the AGNES or DIANA clustering
result.
p = dim(data)[2]
cc = estimated
nc = max(cc)
data.matrix = cbind(data,estimated)
diametros = rep(0,nc)
for (k in 1:nc) {
data.temp = as.matrix(data[data.matrix[,(p+1)]==k,1:p])
maxd = rep(0,(length(which(cc==k))))
for (i in 1:(length(which(cc==k)))) {
maxd[i] = far(data.temp[i,],data.temp) }
diametros[k] = max(maxd,na.rm="TRUE") }
diam.max = max(diametros,na.rm="TRUE")
minimos = matrix(,nc-1,nc)
for (i in 1:(nc-1)) {
data.temp1 = as.matrix(data[data.matrix[,(p+1)]==i,1:p])
```

```
for (k in (i+1):nc) {
data.temp2 = as.matrix(data[data.matrix[,(p+1)]==k,1:p])
mind = rep(0,dim(data.temp1)[1])
for (j in 1:dim(data.temp1)[1]) {
mind[j] = near(data.temp1[j,],data.temp2,diam.max) }
minimos[i,k] = min(mind,na.rm="TRUE") } }
min.t = matrix(,dim(minimos)[1],1)
for (l in 1:dim(minimos)[1]) {
min.t[l] = min(minimos[l,],na.rm="TRUE") }
return(min(min.t,na.rm="TRUE")) } }
```

Davies-Bouldin Program

```
daviesb = function(data,estimated) {
p = dim(data)[2]
cc = estimated
nc = max(cc)
data.matrix = cbind(data,estimated)
Calculating the dispersion of the c(k) cluster
si = rep(0,nc)
for (k in 1:nc) {
data.temp = as.matrix(data[data.matrix[,(p+1)]==k,1:p])
means = apply(as.matrix(data.temp),2,mean)
dev = sweep(data.temp,2,means)
dev = dev^2
suma = apply(dev,1,sum)
prom = mean(suma)
prom1 = sqrt(prom)
si[k] = prom1 }
Calculating the dissimilarity between clusters
l = matrix(,nc,nc)
ri = rep(0,nc)
for (i in 1:nc) {
data.temp1 = as.matrix(data[data.matrix[,(p+1)]==i,1:p])
```

```
data.temp1=as.matrix(data.temp1)
dim1=ncol(data.temp1)
if(dim1>1){means1=apply(data.temp1,2,mean)}
if(dim1==1){means1=apply(t(data.temp1),2,mean)}
for (k in 1:nc) {
data.temp2 = as.matrix(data[data.matrix[,(p+1)]==k,1:p])
data.temp2=as.matrix(data.temp2)
dim2=ncol(data.temp2)
if(dim2>1){means2=apply(data.temp2,2,mean)}
if(dim2==1){means2=apply(t(data.temp2),2,mean)}
dik=distancia(means1,means2)
l[i,k]=ifelse(i==k,NA,(si[i]+si[k])/dik)}
ri[i] = max(l[i,],na.rm=TRUE)}
sum(ri)/nc }
```

Far Function

```
far = function(x, data) {
*****************************************
This function find the maximum distance
between a vector and a matrix.
*****************************************
nd = length(data[,1])
distall = rep(0, nd)
for(i in 1:nd) {
distall[i] = distancia(x, data[i,]) }
maximus = max(distall)
dist.max = maximus
return(dist.max) }
```

Near Function

```
near=function(x, data,maximun) {
*************************************
This function find the minimum distance
between a vector and a matrix.
```

```
*****************************************
nd = length(data[, 1])
distall = rep(0, nd)
for(i in 1:nd) {
distall[i] = distancia(x, data[i,]) }
minimus = min(distall)
dist.min = minimus/maximun
return(dist.min) }
```

Distancia Function

```
distancia = function(x,y) {
************************************
This function find the distance between
two vectors x and y.
****************************************
distancia = 0
for (i in 1:length(x)) {
distancia = distancia+((x[i] − y[i])^2) }
distancia = sqrt(distancia)
distancia }
```

Rand Function

```
rand = function(true.labels, estimated.labels) {
ndat = length(true.labels)
tab = table(true.labels,estimated.labels)
nidot = apply(tab, 1, sum)
ni2 = sum(nidot^2)
ndotj = apply(tab, 2, sum)
nj2 = sum(ndotj^2)
z = sum(tab^2)
num = z-ndat
tempo = (ni2-ndat )* (nj2-ndat)
rdenom = ndat*(ndat−1)
```

```
rand = 1+2*(z-.5*(ni2+nj2))/rdenom rand }
```

Jaccard Function

```
jaccard = function(true.labels, estimated.labels) {
ndat = length(true.labels)
tab = table(true.labels,estimated.labels)
nidot = apply(tab, 1,sum)
ni2 = sum(nidot^2)
ndotj = apply(tab, 2, sum)
nj2 = sum(ndotj^2)
z = sum(tab^2)
num = z-ndat
tempo = (ni2-ndat )* (nj2-ndat)
rdenom = ndat*(ndat-1)
jaccard = num / (ni2+nj2-z-ndat)
jaccard }
```

Folkes and Mallows Function

```
fandm = function(true.labels, estimated.labels) {
ndat = length(true.labels)
tab = table(true.labels, estimated.labels)
nidot = apply(tab, 1, sum)
ni2 = sum(nidot^2)
ndotj = apply(tab, 2, sum)
nj2 = sum(ndotj^2)
z = sum(tab^2)
num = z-ndat
tempo = (ni2-ndat )*(nj2-ndat)
rdenom = ndat*(ndat-1)
fandm = num / sqrt(tempo)
fandm }
```

Hubert Function

```
hubert = function(true.labels, estimated.labels) {
ndat = length(true.labels)
tab = table(true.labels, estimated.labels)
nidot = apply(tab, 1, sum)
ni2 = sum(nidot^2)
ndotj = apply(tab, 2, sum)
nj2 = sum(ndotj^2)
z=sum(tab^2)
num = z-ndat
tempo = (ni2-ndat )*(nj2-ndat)
rdenom = ndat*(ndat-1)
hubert = (rdenom*num-tempo)/sqrt(tempo*(rdenom-ni2+ndat)*(rdenom-nj2+ndat))
hubert }
```

APPENDIX B
DAVIES-BOULDIN INDEX TABLE RESULTS

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
DB_{11}	**0.3183**	0.5674	0.6525	0.6193	0.8226	0.9112	0.7601	0.9869	0.9714
DB_{12}	0.3183	**0.2907**	0.4777	0.4485	0.5397	0.5285	0.4999	0.5898	0.5531
DB_{13}	**0.2920**	0.7950	0.7891	0.7543	0.6427	0.7189	0.8397	1.0526	1.0728
DB_{14}	**0.3183**	0.8697	0.8122	0.9998	1.0342	1.0407	1.0397	1.1761	1.1865
DB_{15}	**0.3634**	0.3739	1.0344	0.8857	0.8906	1.0215	1.2338	1.1756	1.2143
DB_{21}	**0.2977**	0.5476	0.7357	0.6344	0.7758	0.7229	0.8532	0.8459	0.9091
DB_{22}	**0.2977**	0.4523	0.4065	0.4485	0.5397	0.5964	0.5574	0.6238	0.8575
DB_{23}	**0.2918**	0.7320	0.7537	0.7794	1.1350	1.2914	1.1681	1.1837	1.1897
DB_{24}	**0.2977**	0.6941	0.9033	0.8661	0.8734	0.9751	1.0162	1.0411	1.0467
DB_{25}	**0.2977**	0.7397	0.7204	0.7407	0.6606	0.7687	0.8365	0.9116	0.9618
DIANA									
Eucl.	**0.2977**	0.7906	0.7819	0.8475	0.9187	1.0412	0.9475	0.9433	1.0806
Manh.	**0.2977**	0.7197	0.7293	0.9211	0.9454	0.9883	0.9033	1.1302	1.1717
PAM									
Eucl.	**0.2977**	0.8034	0.9836	0.9079	0.9046	1.0052	1.0212	1.0460	1.0776
Manh.	**0.2977**	0.7513	0.9029	0.9593	0.9019	0.9233	0.9913	1.1735	1.2946

Table B–1: DB Index, Iris Data Set

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
DB_{11}	**0.4780**	0.9290	1.4065	1.2338	1.3362	1.6751	1.7125	1.8822	1.7616
DB_{12}	*0.0000*	*0.0000*	*0.0000*	0.3481	0.4802	0.4650	0.4648	0.4784	0.6015
DB_{13}	**0.5830**	1.4472	1.4148	1.8017	1.7916	1.9525	2.0045	1.9891	2.1230
DB_{14}	**0.6469**	1.7269	2.0148	1.9542	1.9209	1.7959	1.7278	1.9206	1.8390
DB_{15}	0.8514	0.9236	**0.8220**	1.1678	1.6998	1.5330	1.5421	1.6495	1.5188
DB_{21}	**0.4586**	0.9471	1.3598	1.1735	1.2119	1.5408	1.5415	1.5512	1.5456
DB_{22}	*0.0000*	*0.0000*	0.3387	0.3364	0.5465	0.5199	0.5070	0.4843	0.5193
DB_{23}	**0.5958**	1.3856	1.8309	1.7549	1.9641	2.0965	2.0105	2.1151	2.1906
DB_{24}	**0.6113**	1.3785	1.8663	2.3912	2.2876	2.0851	1.8955	2.0854	2.0169
DB_{25}	**0.5971**	1.7027	1.5315	1.5801	1.3570	1.5484	1.9199	1.8440	1.7638
DIANA									
Eucl.	**0.5579**	1.6515	1.6179	1.5107	1.7731	1.8179	1.6531	1.6467	1.7532
Manh.	**0.5511**	1.5518	1.8146	1.9277	1.8860	1.9465	1.8591	1.8629	1.8355
PAM									
Eucl.	**0.5919**	1.4912	1.3218	1.6045	1.9277	2.0313	2.1568	2.1210	2.0421
Manh.	**0.5740**	1.7522	1.5068	1.6269	2.0393	1.8791	2.0227	2.0400	1.9554

Table B–2: DB Index values, Breastw Data Set

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
DB_{11}	*0.0000*	*0.0000*	*0.0000*	*0.0000*	0.4913	0.6034	0.7513	0.6383	0.6352
DB_{12}	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*
DB_{13}	3.0736	2.5646	2.1320	**2.1085**	2.5452	2.2940	2.2034	2.1394	2.0242
DB_{14}	**1.0775**	1.6007	1.5959	1.6292	2.1789	2.0728	1.9475	2.0360	2.0016
DB_{15}	3.7032	3.1368	2.5432	2.2870	1.9975	1.8072	1.7493	1.7399	**1.6808**
DB_{21}	0.9579	1.2365	1.4829	1.2138	1.2082	1.0008	**0.9458**	0.9680	1.0852
DB_{22}	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*
DB_{23}	**1.3092**	2.2697	2.2120	2.0195	2.0052	2.1398	1.9363	1.8864	1.8558
DB_{24}	**1.0378**	2.1486	1.6635	2.0343	2.3543	2.2924	2.4526	2.5822	2.4594
DB_{25}	**1.3599**	2.2046	2.1130	1.8177	1.7775	1.9495	1.8211	1.5928	1.4577
DIANA									
Eucl.	1.6870	1.3149	1.1981	1.0169	**0.7306**	0.8531	1.3790	1.2054	1.2868
Manh.	**1.2286**	1.8347	2.0355	1.8177	1.8687	2.2174	2.2799	2.1084	1.9302
PAM									
Eucl.	**1.0553**	1.6646	1.6951	2.0356	1.9959	1.8911	1.8059	1.7555	2.1451
Manh.	**1.0094**	1.9018	1.6920	1.8276	1.7460	1.7334	1.7547	1.7651	2.0369

Table B–3: DB Index values, Ionosphere Data Set

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
DB_{11}	**0.0792**	0.6445	0.6214	0.3762	0.7698	0.7585	0.5690	0.4309	0.9116
DB_{12}	*0.0000*	0.0841	0.2843	*0.0000*	*0.0000*	*0.0000*	0.4184	1.3494	1.3318
DB_{13}	**0.0918**	1.1710	0.8898	0.8108	1.7733	1.8054	1.5996	0.9654	0.8979
DB_{14}	**0.0918**	0.8560	1.1169	1.3585	1.2474	1.2033	1.1878	1.2785	1.0238
DB_{15}	**0.0792**	0.3629	0.2843	0.9397	1.0766	0.7585	0.5690	1.0160	0.9116
DB_{21}	**0.0792**	0.3629	0.8219	0.9230	0.9638	0.8319	0.6523	0.4898	0.9031
DB_{22}	0.0792	0.3629	*0.0000*	*0.0000*	*0.0000*	*0.0000*	0.7137	1.3974	1.4136
DB_{23}	**0.0918**	1.2863	0.9774	1.0638	0.8340	1.3057	1.3012	1.1846	1.0111
DB_{24}	**0.0918**	0.8560	2.3922	1.3155	1.3335	1.2744	1.1985	1.2594	1.0784
DB_{25}	**0.0792**	0.8560	0.9774	1.3155	1.2111	1.2383	1.1985	1.0041	0.9031
DIANA									
Eucl.	**0.0792**	1.7302	0.9774	0.9230	0.7114	0.8319	1.1985	1.0041	1.0584
Manh.	**0.0792**	0.8560	0.9774	1.3155	1.2487	1.2744	1.1985	1.4282	1.4234
PAM									
Eucl.	**0.0863**	2.5218	2.2442	2.2358	2.2970	1.5805	1.6170	1.6269	0.9837
Manh.	**0.0863**	2.5218	2.2442	1.6019	1.5859	1.5655	1.5560	1.6109	1.6221

Table B–4: DB Index values, Breastcc Data Set

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
DB_{11}	**0.0285**	0.0386	0.0394	0.0417	0.0420	0.0419	0.6426	0.5232	0.5183
DB_{12}	**0.0285**	0.0386	0.0393	0.0398	0.0475	0.0469	0.3727	0.3652	0.3634
DB_{14}	0.8224	**0.0386**	0.9818	1.4441	1.9626	1.7970	1.7600	1.6438	1.6080
DB_{14}	**1.2895**	1.8640	2.0293	1.9034	1.7644	1.6729	1.5538	1.4380	1.5400
DB_{15}	**0.0285**	0.0386	1.2807	1.1864	1.0819	1.0541	1.3852	1.3783	1.3306
DB_{21}	**0.0285**	0.0386	0.0416	0.0417	0.0420	0.9996	0.9858	1.1506	1.1214
DB_{22}	**0.0285**	0.0386	0.0499	0.0486	0.0481	0.0498	0.3739	0.3675	0.3681
DB_{23}	0.8718	**0.5838**	1.1509	1.8010	1.8776	1.8336	1.6913	1.6886	1.5670
DB_{24}	**1.2781**	1.8984	1.9331	1.8334	1.7953	1.9340	1.8789	1.7950	1.6552
DB_{25}	**0.0285**	1.2847	1.1277	1.7183	1.5589	1.4771	1.4737	1.3583	1.2012
DIANA									
Eucl.	0.8718	**0.5838**	1.3067	1.5735	1.6069	1.4161	1.5519	1.5604	1.4962
Manh.	0.8718	**0.5838**	1.3067	1.6170	1.6088	1.6968	1.7620	1.5882	1.4657
PAM									
Eucl.	**1.4340**	2.1838	2.0576	1.7004	1.8493	1.9250	1.8285	1.9336	1.8056
Manh.	**1.6110**	2.1475	1.9679	1.8334	1.9598	1.9703	1.9656	1.7853	1.7056

Table B–5: DB Index values, SRBCT Data Set

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
DB_{11}	**0.0164**	0.0233	0.0239	0.0261	0.0273	0.0282	0.0287	0.0291	0.5249
DB_{12}	**0.0179**	0.0256	0.0260	0.0266	0.0273	0.0282	0.0289	0.0292	0.0302
DB_{13}	1.3000	1.5320	1.3138	1.1972	1.1088	1.2490	1.1857	1.1019	**1.0640**
DB_{14}	1.6186	1.8556	1.9820	1.5879	1.9081	1.7948	1.6006	1.4661	**1.3598**
DB_{15}	**0.0164**	0.0233	0.0257	0.0261	0.0273	0.0282	0.6243	0.6193	0.6883
DB_{21}	**0.0179**	0.6670	0.6335	0.8353	0.7925	0.5534	0.5520	0.9476	0.9903
DB_{22}	**0.0179**	0.0256	0.0260	0.0280	0.0283	0.0282	0.0289	0.0294	0.0302
DB_{23}	**1.0183**	1.3208	1.5544	1.3062	1.0521	1.3464	1.4952	1.3081	1.2195
DB_{24}	**1.2660**	1.4737	2.3149	2.1772	2.0037	1.7774	1.6904	1.5247	1.4390
DB_{25}	**0.0179**	0.9700	0.9103	1.1959	0.9316	0.7655	0.5520	1.0130	0.9560
DIANA									
Eucl.	1.2816	1.0828	0.9804	0.9298	0.8827	**0.8641**	1.2743	1.1954	1.1461
Manh.	1.2816	1.0859	1.0010	1.1999	1.1072	**1.0558**	1.3643	1.4007	1.3010
PAM									
Eucl.	1.9300	1.7194	1.8774	2.0538	1.8265	1.6442	1.4922	1.3616	**1.2838**
Manh.	1.3521	1.7285	1.6816	2.0905	1.8362	1.6421	1.4800	1.3587	**1.2952**

Table B–6: DB Index values, Brain Data Set

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
DB_{11}	**0.0222**	0.6309	0.5877	0.5752	0.8350	0.6548	0.8762	1.3369	1.2797
DB_{12}	**0.0222**	0.0329	0.0338	0.0348	0.0375	0.0383	0.0406	0.4357	0.4318
DB_{13}	1.7136	**1.3889**	2.0047	2.0911	2.0581	1.9850	2.0708	1.9120	1.8713
DB_{14}	**1.5693**	2.4034	2.4857	2.3602	2.3567	2.0670	2.1518	2.0340	1.9800
DB_{15}	**0.0222**	1.3638	1.3903	1.2733	1.3581	1.3784	1.4634	1.2979	1.1547
DB_{21}	**0.0222**	0.0294	0.0332	0.0373	0.0382	0.9776	1.2276	1.2484	1.2219
DB_{22}	**0.0222**	0.0308	0.0338	0.0358	0.0382	0.4992	0.5005	0.4937	0.4936
DB_{23}	1.7227	**1.3947**	2.0140	1.7647	1.7251	1.9163	1.9429	1.9869	2.0082
DB_{24}	**1.7619**	2.3511	2.4240	2.3437	2.2468	1.9802	2.0615	1.9966	2.0006
DB_{25}	**0.0222**	1.2770	1.7598	1.5417	1.3407	1.2227	1.1573	1.5365	1.6492
DIANA									
Eucl.	**0.0222**	1.2747	1.6211	1.4612	1.3466	1.5147	1.6052	1.6024	1.6206
Manh.	**0.0222**	1.2903	1.8063	1.7693	1.6074	1.9209	1.7736	1.8491	1.8695
PAM									
Eucl.	**1.5976**	2.1654	2.0506	2.6497	2.4039	2.6267	2.3708	2.0779	2.1917
Manh.	**1.6083**	2.2003	2.6306	3.1604	2.4300	2.5479	2.5134	2.3856	2.1367

Table B–7: DB Index values, Leukemia Data Set

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
DB_{11}	0.7927	**0.7232**	1.3128	1.2034	1.1124	1.0659	1.3745	1.4230	1.3441
DB_{12}	**0.0153**	0.0196	0.0210	0.6658	0.6507	0.6402	1.0382	0.9966	0.9650
DB_{13}	**1.2396**	1.9020	2.2117	2.1549	2.2338	2.0131	1.9626	1.8874	1.7241
DB_{14}	**0.7927**	1.3916	1.6609	1.8510	1.9585	1.8312	1.8011	1.9109	1.8758
DB_{15}	0.7927	**0.7232**	1.3921	1.2523	1.1743	1.4533	1.4928	1.5182	1.3858
DB_{21}	0.7927	**0.7232**	1.3358	1.2058	1.6194	1.7155	1.6489	1.5866	1.5032
DB_{22}	**0.0153**	0.7232	0.6840	0.6658	1.0883	1.0454	1.0231	0.9856	0.9888
DB_{23}	**0.7607**	1.5376	2.3153	2.1163	1.9188	1.9687	1.8348	1.8771	1.8877
DB_{24}	**0.7927**	1.3916	1.6082	2.1043	2.0825	1.9917	1.9415	1.9583	1.8731
DB_{25}	0.7927	**0.7458**	1.3317	1.2048	1.6433	1.5321	1.6643	1.6677	1.6182
DIANA									
Eucl.	**0.8352**	1.5616	1.6328	1.4414	1.8219	1.7119	1.5615	1.7125	1.6614
Manh.	**0.7927**	1.5040	1.2833	1.1599	1.6163	1.5607	1.6830	1.6751	1.6617
PAM									
Eucl.	**0.8795**	1.5102	1.7800	1.9804	2.1721	2.1899	2.0398	1.8612	1.8224
Manh.	**0.8352**	1.5102	1.8204	1.9544	2.1769	2.1652	2.0933	1.9822	1.7867

Table B–8: DB Index values, Lymphoma Data Set

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
DB_{11}	**0.0113**	0.5912	0.7070	0.7341	0.7410	0.7715	0.9765	0.9290	0.8964
DB_{12}	**0.0113**	0.0234	0.0276	0.0307	0.0325	0.0340	0.0346	0.0363	0.6797
DB_{13}	**0.0113**	0.5713	0.6966	0.7270	0.7473	0.9989	0.9661	0.9943	0.8811
DB_{14}	**0.5008**	0.7547	0.7020	1.3004	1.2199	1.4410	1.5664	1.3577	1.4418
DB_{15}	**0.0113**	0.5713	0.6731	0.7083	0.7180	0.7315	0.7672	0.9055	0.8763
DB_{21}	**0.0113**	0.5713	0.7069	0.7327	0.7521	0.7546	0.7516	0.7548	0.7337
DB_{22}	**0.0113**	0.0234	0.0294	0.0307	0.0325	0.0328	0.7081	0.7095	0.7072
DB_{23}	0.5015	**0.4338**	0.7284	1.3293	1.1813	1.5224	1.6515	1.5459	1.4569
DB_{24}	**0.5015**	0.7836	1.5195	1.3372	1.6640	1.8239	1.8245	1.8605	1.7230
DB_{25}	**0.0113**	0.4633	0.4358	0.7503	0.8596	1.1285	0.9074	0.8794	0.9618
DIANA									
Eucl.	**0.0113**	0.5821	0.7485	0.7663	0.7803	0.7804	0.7921	1.1341	0.9595
Manh.	**0.5964**	0.6269	0.8190	0.8294	1.1956	1.1175	1.1443	1.3916	1.3329
PAM									
Eucl.	**0.5015**	0.6678	1.2786	1.1357	1.4291	1.3295	1.2603	1.2077	1.1614
Manh.	**0.5015**	0.7410	1.4431	1.2728	1.5036	1.8775	1.6530	1.5564	1.6884

Table B–9: DB Index values, Prostate Data Set

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
DB_{11}	1.0015	0.7790	**0.4776**	0.7946	0.9355	1.2508	1.1310	1.1152	1.0685
DB_{12}	**0.0332**	0.0517	0.0543	0.4578	0.4509	0.4459	0.4492	0.4508	0.4443
DB_{13}	**1.2994**	1.9167	2.1574	1.8302	1.8330	1.7699	1.7451	1.6601	1.5984
DB_{14}	**1.4792**	1.9639	1.6998	1.9977	1.8867	1.8450	1.8155	1.8123	1.7676
DB_{15}	**0.9075**	1.6703	1.8614	1.6242	1.5869	1.5278	1.3733	1.2948	1.3220
DB_{21}	1.0015	**0.7790**	1.0605	0.7946	0.9355	1.2508	1.1310	1.0619	1.0558
DB_{22}	**0.0332**	0.0496	0.4855	0.4717	0.4509	0.4459	0.4492	0.4612	0.4561
DB_{23}	**1.2994**	1.9067	2.1611	1.9750	1.9931	1.9116	1.8507	1.8486	1.7858
DB_{24}	**1.3708**	2.2387	2.0145	2.0365	1.9103	1.9650	1.8856	1.8368	1.6869
DB_{25}	**0.7638**	0.9524	1.6555	1.3898	1.4566	1.4718	1.3701	1.2369	1.4376
DIANA									
Eucl.	**1.0015**	1.4829	1.2933	1.8823	1.7836	1.7145	1.5551	1.5450	1.7248
Manh.	**1.5063**	2.0360	2.0639	1.8531	1.7494	1.7056	1.9049	1.7931	1.8466
PAM									
Eucl.	**1.7015**	2.4681	2.1741	2.1136	2.1379	1.9780	2.0518	1.9868	1.7995
Manh.	**1.6921**	2.6119	2.1741	2.0086	2.1049	2.1045	2.0566	1.9899	1.8651

Table B–10: DB Index values, Colon Data Set

APPENDIX C
DUNN INDEX TABLE RESULTS

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
D_{11}	0.0924	0.0924	**0.1504**	0.0834	0.0783	0.0930	0.0930	0.0930	0.0537
D_{12}	**0.3389**	0.1307	0.0654	0.0654	0.0944	0.0944	0.0629	0.0629	0.0678
D_{13}	0.0924	**0.1290**	0.1290	0.1290	0.1290	0.0745	0.0868	0.0868	0.0922
D_{14}	**0.3389**	0.1691	0.1290	0.1290	0.1290	0.1012	0.0745	0.0745	0.0745
D_{15}	0.0304	0.0339	0.0339	0.0339	0.0481	0.0481	0.0556	0.0629	**0.0710**
D_{21}	0.0373	0.0428	0.0618	0.0505	0.0556	**0.0629**	0.0629	0.0629	0.0629
D_{22}	**0.0924**	0.0270	0.0270	0.0339	0.0339	0.0339	0.0339	0.0339	0.0594
D_{23}	**0.3389**	0.0909	0.0681	0.0556	0.0556	0.0629	0.0629	0.0629	0.0760
D_{24}	0.0310	0.0310	0.0428	0.0428	0.0428	0.0525	**0.0618**	0.0618	0.0618
D_{25}	**0.3389**	0.0654	0.0552	0.0798	0.0798	0.0798	0.0880	0.0880	0.1194
DIANA									
Eucl.	**0.3389**	0.0678	0.0429	0.0429	0.0429	0.0556	0.0556	0.0556	0.0629
Manh.	**0.3389**	0.0487	0.0487	0.0487	0.0629	0.0629	0.0629	0.0629	0.0629
PAM									
Eucl.	**0.3389**	0.0299	0.0415	0.0415	0.0668	0.0679	0.0679	0.0624	0.0770
Manh.	**0.3389**	0.0452	0.0679	0.0582	0.0620	0.0620	0.0758	0.1199	0.0758

Table C–1: Dunn Index values, Iris Data Set

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
D_{11}	0.1400	0.1320	0.1402	0.1402	0.1638	0.1638	**0.1727**	0.1727	0.1727
D_{12}	0.1504	0.1504	0.2016	0.2016	0.2016	0.2016	0.2016	0.1948	**0.2075**
D_{13}	**0.3559**	0.3517	0.2487	0.2162	0.2162	0.2162	0.2162	0.2162	0.2162
D_{14}	**0.3559**	0.2298	0.2298	0.2265	0.2162	0.2162	0.2162	0.2162	0.2162
D_{15}	0.1545	0.1638	0.1638	0.1679	**0.1912**	0.1912	0.1912	0.1876	0.1876
D_{21}	0.1291	0.1410	0.1533	0.1533	0.1615	0.1615	0.1615	0.1768	**0.1853**
D_{22}	0.1228	0.1316	0.1316	0.1316	**0.1518**	0.0480	0.0480	0.0482	0.0482
D_{23}	0.0964	0.1041	0.1114	0.1114	0.1237	0.1252	0.1252	0.1252	**0.1300**
D_{24}	0.1320	0.1320	0.1320	0.1446	0.1446	0.1446	**0.1541**	0.1541	0.1541
D_{25}	0.1504	0.1781	0.1781	0.1781	0.1781	0.1809	**0.1848**	0.1848	0.1848
DIANA									
Eucl.	0.1545	0.1622	0.1622	0.1653	0.1674	0.1725	0.1725	**0.1773**	0.1630
Manh.	0.0975	0.1045	0.1091	0.1091	0.1091	0.1091	**0.1197**	0.1197	0.1197
PAM									
Eucl.	0.1430	**0.1693**	0.0510	0.0513	0.0513	0.0513	0.0513	0.0528	0.0528
Manh.	**0.1339**	0.1289	0.0487	0.0546	0.0563	0.0563	0.0563	0.0563	0.0578

Table C–2: Dunn Index values, Breastw Data Set

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
D_{11}	**0.5706**	0.5373	0.5168	0.5300	0.4824	0.4824	0.4472	0.4472	0.4472
D_{12}	**0.4634**	0.4204	0.4472	0.4472	0.4472	0.4472	0.4472	0.4472	0.3993
D_{13}	**0.5706**	0.5584	0.5510	0.5300	0.5187	0.5187	0.5104	0.5104	0.5027
D_{14}	**0.5706**	0.5168	0.5058	0.4902	0.4902	0.4584	0.4584	0.4584	0.4584
D_{15}	0.0644	0.0679	0.0679	0.0724	0.0724	0.0759	0.0759	0.0813	**0.0816**
D_{21}	0.1155	**0.1248**	0.0677	0.0677	0.0691	0.0729	0.0742	0.0742	0.0742
D_{22}	0.0719	0.0719	0.0719	0.0719	**0.0763**	0.0597	0.0597	0.0671	0.0671
D_{23}	0.0780	0.0780	0.0780	0.0780	**0.0805**	0.0630	0.0630	0.0630	0.0630
D_{24}	0.0634	0.0677	0.0677	**0.0708**	0.0708	0.0708	0.0708	0.0708	0.0708
D_{25}	0.3641	0.0826	0.0856	0.0856	**0.0890**	0.0890	0.0890	0.0890	0.0890
DIANA									
Eucl.	0.2701	0.2701	0.2701	0.2701	0.3004	0.3004	0.3004	0.3004	**0.3066**
Manh.	**0.4204**	0.1627	0.1627	0.1726	0.1726	0.0852	0.0852	0.0879	0.0954
PAM									
Eucl.	**0.0718**	0.0350	0.0573	0.0505	0.0423	0.0458	0.0380	0.0380	0.0411
Manh.	**0.0698**	0.0308	0.0621	0.0505	0.0251	0.0434	0.0400	0.0400	0.0400

Table C–3: Dunn Index values, Ionosphere Data Set

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
D_{11}	0.7854	0.7853	0.7327	0.7582	0.7689	0.7689	0.7737	**0.8144**	0.8144
D_{12}	**0.7854**	0.7853	0.7571	0.7551	0.7109	0.7085	0.6970	0.6944	0.7344
D_{13}	0.6952	0.7387	0.7514	0.7139	0.7198	0.7621	0.8093	0.8158	**0.8195**
D_{14}	0.6952	0.7169	0.6999	0.7345	0.7345	0.7345	0.7810	**0.8275**	0.8275
D_{15}	0.7854	0.7139	0.7365	0.7365	0.7365	0.7737	**0.8240**	0.8240	0.8240
D_{21}	0.7854	0.7853	0.7327	0.7582	0.7619	0.7737	0.7737	**0.8144**	0.8144
D_{22}	**0.7854**	0.7853	0.7571	0.7551	0.7085	0.6970	0.6970	0.6944	0.7689
D_{23}	0.6952	0.7387	0.7514	0.7139	0.7198	0.7592	0.8093	0.8180	**0.8195**
D_{24}	0.6952	0.7169	0.6999	0.7345	0.7345	0.7904	0.8144	**0.8191**	0.8191
D_{25}	0.7854	0.7169	0.7514	0.7345	0.7345	0.7345	**0.8144**	0.8144	0.8144
DIANA									
Eucl.	0.7854	0.7169	0.7514	0.7139	0.7198	0.7621	0.7916	0.8093	**0.8195**
Manh.	0.7854	0.7169	0.7034	0.7345	0.7389	0.7904	0.8062	**0.8144**	0.8066
PAM									
Eucl.	0.6589	0.6654	0.6999	0.6999	0.7064	0.7220	0.7220	**0.7389**	0.7389
Manh.	0.6589	0.6654	0.6999	0.7064	0.7064	0.7220	0.7220	0.7220	**0.7389**

Table C–4: Dunn Index values, Breastcc Data Set

APPENDIX D
SILHOUETTE INDEX TABLES RESULTS

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
Si_{11}	**0.6730**	0.4755	0.3910	0.2971	0.2737	0.2914	0.2464	0.2046	0.2055
Si_{12}	**0.6730**	0.5590	0.3658	0.2938	0.1156	0.1109	0.1448	0.0774	0.0908
Si_{13}	**0.5226**	0.4730	0.4454	0.3192	0.3046	0.2878	0.2850	0.2377	0.2400
Si_{14}	**0.6730**	0.4638	0.3259	0.2465	0.2447	0.2363	0.2017	0.1870	0.1943
Si_{15}	**0.3572**	0.0964	0.2351	0.2237	0.1709	0.1988	0.2015	0.1670	0.1470
Si_{21}	**0.6864**	0.4921	0.3830	0.2780	0.3080	0.2602	0.2501	0.2149	0.2023
Si_{22}	**0.6864**	0.5118	0.3853	0.2938	0.1156	0.0502	−0.0002	−0.0544	−0.0802
Si_{23}	**0.5366**	0.4605	0.4462	0.3040	0.2317	0.2362	0.2311	0.2210	0.2059
Si_{24}	**0.6864**	0.5330	0.4472	0.3246	0.3045	0.2958	0.2932	0.2504	0.2473
Si_{25}	**0.6864**	0.4558	0.3259	0.3113	0.3003	0.2752	0.2981	0.2731	0.2650
DIANA									
Eucl.	**0.6864**	0.4973	0.3498	0.3054	0.2932	0.2498	0.2445	0.2222	0.2108
Manh.	**0.6864**	0.5280	0.3880	0.3048	0.2976	0.2889	0.2838	0.2390	0.2347
PAM									
Eucl.	**0.5802**	0.4555	0.4080	0.3620	0.3422	0.3287	0.3345	0.3333	0.3531
Manh.	**0.6192**	0.4876	0.4367	0.4226	0.3418	0.3344	0.3390	0.3071	0.3212

Table D–1: Silhouette values obtained from Iris Samples

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
Si_{11}	**0.4333**	0.1519	0.4501	0.4472	0.4437	0.4260	0.4250	0.4027	0.4026
Si_{12}	**0.4073**	0.3829	0.3311	0.2993	0.2689	0.1620	0.1621	0.1586	0.1610
Si_{13}	**0.5499**	0.5011	0.3702	0.3522	0.3887	0.3841	0.4237	0.4252	0.4242
Si_{14}	**0.5683**	0.5261	0.4859	0.4802	0.4795	0.2076	0.2088	0.2086	0.2126
Si_{15}	**0.3865**	0.1273	0.0750	0.3314	0.2987	0.2989	0.2943	0.2864	0.1856
Si_{21}	0.4335	0.2741	**0.5069**	0.4756	0.4707	0.4501	0.4490	0.4432	0.4652
Si_{22}	**0.4073**	0.2401	0.1858	0.1483	0.1075	0.1089	0.1097	0.1125	0.1164
Si_{23}	**0.5867**	0.5359	0.5148	0.4364	0.4207	0.4144	0.4146	0.4175	0.4196
Si_{24}	**0.5837**	0.5323	0.4860	0.4749	0.4398	0.4432	0.4463	0.4457	0.4454
Si_{25}	0.4116	**0.5371**	0.5056	0.4997	0.4992	0.4931	0.4890	0.4827	0.4824
DIANA									
Eucl.	**0.5558**	0.5056	0.3079	0.3006	0.2969	0.4113	0.4115	0.4077	0.4054
Manh.	**0.5698**	0.5160	0.4968	0.4790	0.4528	0.4521	0.4568	0.4487	0.4461
PAM									
Eucl.	**0.5717**	0.5501	0.2366	0.1711	0.1648	0.1665	0.1619	0.1683	0.1636
Manh.	**0.6408**	0.5781	0.1947	0.2021	0.1989	0.1685	0.1885	0.1981	0.1989

Table D–2: Silhouette values obtained from Breastw Samples

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
Si_{11}	**0.4050**	0.3613	0.3458	0.3306	0.3301	0.3278	0.3298	0.3313	0.3260
Si_{12}	**0.4050**	0.3508	0.3412	0.3306	0.3284	0.3280	0.3041	0.3035	0.2953
Si_{13}	0.1574	0.1983	0.1366	0.1468	0.2023	0.2213	0.2243	0.2278	**0.2285**
Si_{14}	0.2942	0.2965	**0.3066**	0.2449	0.2532	0.2043	0.2061	0.2127	0.2101
Si_{15}	**0.1947**	0.1708	0.1750	0.1772	0.1781	0.1793	0.1819	0.1822	0.1771
Si_{21}	**0.3682**	0.3494	0.3399	0.3360	0.3330	0.3266	0.3014	0.2785	0.2731
Si_{22}	**0.4050**	0.3688	0.3518	0.2999	0.3007	0.3027	0.3023	0.3028	0.2912
Si_{23}	0.2779	**0.2800**	0.1401	0.1573	0.1607	0.1624	0.1641	0.1699	0.1711
Si_{24}	0.2929	0.2921	**0.2975**	0.2802	0.2846	0.2130	0.2208	0.2253	0.2298
Si_{25}	**0.3342**	0.2869	0.2873	0.2889	0.2904	0.2891	0.2899	0.2962	0.2992
DIANA									
Eucl.	**0.3617**	0.3409	0.3349	0.3327	0.3322	0.3309	0.2981	0.2672	0.2703
Manh.	**0.3720**	0.2436	0.2339	0.2323	0.2199	0.2487	0.2516	0.2441	0.2415
PAM									
Eucl.	0.2714	**0.2834**	0.2739	0.1477	0.1540	0.1623	0.1663	0.1579	0.1584
Manh.	0.2808	0.2866	**0.3108**	0.1390	0.1507	0.1593	0.1659	0.1638	0.1795

Table D–3: Silhouette values obtained from Ionosphere Samples

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
Si_{11}	−0.6278	0.0461	0.0021	0.0914	0.1108	0.1517	0.1362	0.2156	**0.2167**
Si_{12}	**0.8840**	−0.5264	−0.5348	−0.4578	0.1370	0.1919	0.1384	0.1373	0.1816
Si_{13}	−0.4738	−0.4570	−0.3853	−0.4017	−0.1161	−0.1238	−0.0388	0.1147	**0.1777**
Si_{14}	−0.4738	−0.5253	0.0196	−0.0433	−0.0026	0.0239	0.0115	−0.0119	**0.1258**
Si_{15}	−0.6278	−0.6305	−0.5348	0.0014	0.0208	0.1517	0.1362	0.1364	**0.2167**
Si_{21}	−0.6278	−0.6305	−0.4765	0.0414	0.0216	0.0113	0.1006	0.1826	**0.1879**
Si_{22}	−0.6278	−0.6305	−0.5508	−0.4578	0.1370	0.1919	**0.1961**	0.1258	0.1790
Si_{23}	−0.4738	0.0022	0.0613	−0.0099	−0.0271	−0.0135	−0.0114	0.0556	**0.1448**
Si_{24}	−0.4738	−0.5253	−0.3098	0.0053	−0.0296	−0.0291	0.0172	−0.0209	**0.0684**
Si_{25}	−0.6278	−0.5253	0.0613	0.0053	0.0302	0.0087	0.0172	0.1065	**0.1879**
DIANA									
Eucl.	−0.6278	−0.1219	0.0613	0.0414	0.0057	0.0113	0.0172	**0.1065**	0.0743
Manh.	−0.6278	−0.5253	**0.0613**	0.0053	−0.0344	−0.0291	0.0172	0.0008	−0.0005
PAM									
Eucl.	**0.2370**	0.1160	0.1393	0.1044	0.0760	0.0801	0.0744	0.0309	0.0343
Manh.	**0.2826**	0.2211	0.1425	0.1264	0.0977	0.1022	0.0510	0.0482	0.0203

Table D–4: Silhouette values obtained from Breastcc Samples

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
Si_{11}	**0.0822**	0.0803	0.0674	0.0043	0.0228	0.0273	0.0431	0.0582	0.0629
Si_{12}	0.0822	0.0803	0.0928	0.0762	0.0780	0.0805	0.0948	**0.1139**	0.0807
Si_{13}	0.0818	0.0803	0.0700	0.0727	0.0869	0.0938	0.0751	0.0941	**0.1115**
Si_{14}	0.1132	0.0985	0.1002	0.0993	0.0976	0.1144	0.1243	0.1389	**0.1401**
Si_{15}	0.0822	0.0803	0.0533	0.0510	0.0693	0.0508	0.0695	0.0775	**0.0876**
Si_{21}	0.0822	0.0803	0.0121	0.0043	0.0228	0.0480	0.0532	0.1024	**0.1061**
Si_{22}	**0.0822**	0.0803	0.0525	0.0710	0.0318	0.0307	0.0440	0.0441	0.0488
Si_{23}	0.0685	0.0553	**0.0763**	0.0279	0.0530	0.0459	0.0645	0.0696	0.0752
Si_{24}	0.1095	0.1163	0.1014	0.1039	0.1006	0.0892	0.1025	0.1048	**0.1171**
Si_{25}	**0.0822**	0.0666	0.0836	0.0562	0.0642	0.0540	0.0554	0.0640	0.0814
DIANA									
Eucl.	0.0685	0.0553	0.0881	0.0689	0.0795	0.0953	0.0904	**0.1038**	0.0952
Manh.	0.0685	0.0553	0.0881	**0.0959**	0.0795	0.0691	0.0540	0.0643	0.0799
PAM									
Eucl.	0.0706	0.0793	0.0794	**0.1032**	0.0892	0.0701	0.0737	0.0733	0.0776
Manh.	0.0866	0.1095	0.1074	**0.1262**	0.1033	0.0867	0.0908	0.0981	0.1023

Table D–5: Silhouette values obtained from Srbct Samples

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
Si_{11}	0.1733	0.1411	0.1418	0.1459	0.1685	0.1639	0.1750	**0.1975**	0.1827
Si_{12}	0.1323	0.1127	0.1185	0.1429	0.1685	0.1639	0.1835	0.1882	**0.2071**
Si_{13}	0.1028	0.0982	0.1165	0.1401	0.1570	0.1598	0.1807	0.2065	**0.2206**
Si_{14}	0.0690	0.0482	0.0628	0.0953	0.1050	0.1082	0.1350	0.1604	**0.1848**
Si_{15}	0.1733	0.1411	0.1291	0.1459	0.1685	0.1639	0.1743	**0.1889**	0.1739
Si_{21}	0.1323	0.1040	0.1087	0.1235	0.1196	0.1630	0.1757	**0.2092**	0.1956
Si_{22}	0.1323	0.1127	0.1185	0.1128	0.1391	0.1639	0.1835	0.1803	**0.2071**
Si_{23}	0.0835	0.0771	0.0737	0.1007	0.1225	0.1272	0.1267	0.1745	**0.2008**
Si_{24}	0.1061	0.0588	0.0609	0.0729	0.0782	0.1034	0.1058	0.1325	**0.1330**
Si_{25}	0.1323	0.1147	0.1137	0.0847	0.0942	0.1375	0.1757	0.2044	**0.2274**
DIANA									
Eucl.	0.1062	0.1164	0.1318	0.1516	0.1682	0.1756	0.1842	0.2096	**0.2234**
Manh.	0.1062	0.1193	0.1374	0.1398	0.1570	0.1666	0.1769	0.1601	**0.1863**
PAM									
Eucl.	0.0536	0.0731	**0.0796**	0.0714	0.0753	0.0769	0.0780	0.0783	0.0787
Manh.	**0.1226**	0.0975	0.1033	0.0915	0.0968	0.0995	0.0998	0.1012	0.1026

Table D–6: Silhouette values obtained from Brain Samples

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
Si_{11}	**0.1723**	0.0866	0.0433	0.0456	0.0394	0.0552	0.0578	0.0822	0.0902
Si_{12}	**0.1723**	0.0686	0.0553	0.0442	0.0474	0.0608	0.0637	0.0716	0.0806
Si_{13}	0.0620	0.0737	0.0514	0.0587	0.0698	0.0682	0.0716	0.0850	**0.0886**
Si_{14}	0.0740	0.0453	0.0576	0.0727	0.0698	0.0816	0.0812	0.0851	**0.0875**
Si_{15}	**0.1723**	0.0804	0.0737	0.0623	0.0521	0.0541	0.0607	0.0826	0.0841
Si_{21}	**0.1723**	0.1121	0.0916	0.0847	0.0663	0.0790	0.0778	0.0790	0.0765
Si_{22}	**0.1723**	0.0972	0.0553	0.0655	0.0663	0.0532	0.0593	0.0513	0.0602
Si_{23}	0.0619	**0.0738**	0.0391	0.0467	0.0455	0.0517	0.0634	0.0707	0.0679
Si_{24}	0.0586	0.0307	0.0371	0.0577	0.0676	**0.0837**	0.0742	0.0778	0.0790
Si_{25}	**0.1723**	0.0714	0.0548	0.0591	0.0577	0.0672	0.0732	0.0836	0.1066
DIANA									
Eucl.	**0.1723**	0.0892	0.0757	0.0734	0.0819	0.0779	0.0853	0.0924	0.1006
Manh.	**0.1723**	0.0821	0.0876	0.0789	0.0856	0.0797	0.0921	0.0934	0.0891
PAM									
Eucl.	**0.0734**	0.0486	0.0497	0.0431	0.0420	0.0463	0.0518	0.0548	0.0561
Manh.	**0.0889**	0.0509	0.0544	0.0480	0.0552	0.0629	0.0590	0.0593	0.0623

Table D–7: Silhouette values obtained from Leukemia Samples

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
Si_{11}	**0.2102**	0.1619	0.1410	0.1440	0.1443	0.1469	0.1536	0.1589	0.1655
Si_{12}	0.0452	0.0153	0.0085	0.1288	0.1302	0.1485	0.1572	0.1857	**0.1884**
Si_{13}	0.1218	0.0775	0.0786	0.0908	0.0930	0.1067	0.1118	0.1160	**0.1358**
Si_{14}	**0.2102**	0.1751	0.1144	0.0953	0.1008	0.1031	0.1049	0.1110	0.1145
Si_{15}	**0.2102**	0.1619	0.1287	0.1288	0.1376	0.1356	0.1189	0.1280	0.1389
Si_{21}	**0.2102**	0.1619	0.1355	0.1347	0.1310	0.1344	0.1329	0.1332	0.1415
Si_{22}	0.0452	0.1619	0.1276	0.1288	0.1471	0.1621	0.1671	**0.1698**	0.1630
Si_{23}	**0.2121**	0.1413	0.1247	0.1218	0.1229	0.0798	0.1176	0.1225	0.1194
Si_{24}	**0.2102**	0.1751	0.1258	0.1145	0.1023	0.1042	0.1060	0.1106	0.1111
Si_{25}	**0.2102**	0.1459	0.1191	0.1348	0.1286	0.1264	0.1316	0.1148	0.1199
GS_u	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	c=10
DIANA									
Eucl.	**0.2030**	0.1351	0.1222	0.1203	0.1153	0.1065	0.1175	0.1265	0.1239
Manh.	**0.2102**	0.1429	0.1408	0.1405	0.1379	0.1311	0.1359	0.1548	0.1581
PAM									
Eucl.	**0.1629**	0.1187	0.0860	0.0914	0.0973	0.0924	0.0931	0.0960	0.0985
Manh.	**0.1890**	0.1351	0.0920	0.1022	0.1071	0.1011	0.1036	0.1070	0.1114

Table D–8: Silhouette values obtained from Lymphoma Samples

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
Si_{11}	0.3593	**0.3696**	0.3529	0.3525	0.3566	0.3600	0.3219	0.3223	0.2561
Si_{12}	**0.3593**	0.2873	0.3001	0.2861	0.3032	0.2928	0.2980	0.3150	0.3564
Si_{13}	0.3593	**0.3704**	0.2851	0.2784	0.2811	0.2783	0.2847	0.2821	0.2989
Si_{14}	**0.3327**	0.2812	0.2842	0.1428	0.1364	0.1297	0.1130	0.1312	0.1302
Si_{15}	0.3593	**0.3704**	0.3581	0.3580	0.3620	0.3662	0.2987	0.2959	0.2935
Si_{21}	0.3593	**0.3704**	0.2847	0.2785	0.2812	0.2892	0.2770	0.2851	0.2929
Si_{22}	**0.3593**	0.2873	0.2707	0.2861	0.3032	0.1660	0.2427	0.2537	0.2631
Si_{23}	**0.3362**	0.3069	0.2801	0.2598	0.2671	0.1318	0.1226	0.1229	0.1290
Si_{24}	**0.3362**	0.2771	0.1389	0.1423	0.1229	0.1162	0.0875	0.0836	0.0935
Si_{25}	**0.3593**	0.2979	0.2752	0.2431	0.2693	0.2617	0.2701	0.2759	0.2755
DIANA									
Eucl.	0.3593	**0.3686**	0.2821	0.2763	0.2792	0.2874	0.2936	0.2917	0.2943
Manh.	**0.3815**	0.3588	0.2387	0.2287	0.2311	0.2358	0.2317	0.1330	0.1401
PAM									
Eucl.	**0.2907**	0.1499	0.1040	0.1082	0.0962	0.0967	0.0678	0.0655	0.0741
Manh.	**0.3877**	0.2236	0.1805	0.1817	0.1554	0.1327	0.1198	0.0919	0.0862

Table D–9: Silhouette values obtained from Prostate Samples

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES									
Si_{11}	**0.1359**	0.1150	0.0931	0.0681	0.0781	0.1077	0.1173	0.1150	0.1207
Si_{12}	0.1620	**0.1627**	0.0763	0.0420	0.0524	0.0568	0.0623	0.0731	0.0802
Si_{13}	0.0923	0.0891	0.0621	0.0640	0.0640	0.0631	0.0808	0.0892	**0.0942**
Si_{14}	0.0719	0.0809	**0.0935**	0.0920	0.0887	0.0899	0.0898	0.0884	0.0797
Si_{15}	0.0835	0.0849	0.0999	0.1115	0.1126	0.1185	0.1306	0.1329	**0.1469**
Si_{21}	**0.1359**	0.1150	0.0673	0.0681	0.0781	0.1077	0.1173	0.1184	0.1192
Si_{22}	**0.1620**	0.0597	0.0235	0.0314	0.0524	0.0568	0.0623	0.0662	0.0782
Si_{23}	0.0923	**0.0938**	0.0636	0.0651	0.0596	0.0809	0.0703	0.0733	0.0759
Si_{24}	0.0938	0.0747	0.0812	0.0830	0.0907	0.0937	**0.0996**	0.0943	0.0939
Si_{25}	0.0780	0.0518	0.0705	0.0749	0.0742	0.0689	0.0719	0.0836	**0.1060**
DIANA									
Eucl.	**0.1359**	0.0821	0.0925	0.0758	0.0916	0.0835	0.0996	0.0886	0.0891
Manh.	0.0800	0.0729	0.0812	**0.0828**	0.0611	0.0556	0.0486	0.0522	0.0643
PAM									
Eucl.	**0.0773**	0.0613	0.0749	0.0757	0.0718	0.0763	0.0706	0.0719	0.0740
Manh.	0.0694	0.0703	0.0836	0.0869	0.0866	0.0766	0.0818	0.0810	**0.0890**

Table D–10: Silhouette values obtained from Colon Samples

APPENDIX E
EXTERNAL CRITERIA INDEX RESULTS

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES - Euclidean Distance									
R_{11}	0.7630	0.7640	0.7727	0.7566	**0.7942**	0.7849	0.7852	0.7820	0.7385
R_{12}	0.7630	0.7719	0.7727	0.7725	**0.7730**	0.7726	0.7726	0.7727	0.7729
R_{13}	0.7057	**0.8021**	0.7941	0.7522	0.7543	0.7527	0.7680	0.7699	0.7656
R_{14}	0.7630	**0.8227**	0.7809	0.7719	0.7590	0.7574	0.7284	0.7471	0.7440
R_{15}	0.6209	0.6198	0.7390	0.7357	0.7344	0.7376	0.7470	0.7458	**0.7474**
J_{11}	**0.5778**	0.5693	0.5651	0.5343	0.4946	0.4604	0.4608	0.4440	0.3330
J_{12}	0.5778	**0.5872**	0.5784	0.5781	0.5752	0.5714	0.5612	0.5579	0.5547
J_{13}	0.4745	**0.5692**	0.5276	0.4315	0.4336	0.4299	0.3822	0.3571	0.3416
J_{14}	0.5778	**0.5780**	0.4783	0.4138	0.3689	0.3647	0.2887	0.2944	0.2812
J_{15}	0.3762	0.3280	**0.4155**	0.4083	0.3951	0.3701	0.3634	0.3601	0.3586
FM_{11}	**0.7580**	0.7464	0.7364	0.7080	0.6640	0.6359	0.6363	0.6231	0.5172
FM_{12}	0.7580	**0.7642**	0.7526	0.7523	0.7485	0.7440	0.7319	0.7279	0.7241
FM_{13}	0.6574	**0.7282**	0.6908	0.6038	0.6059	0.6024	0.5738	0.5628	0.5499
FM_{14}	**0.7580**	0.7326	0.6482	0.5969	0.5579	0.5538	0.4758	0.5007	0.4892
FM_{15}	0.5597	0.4979	**0.5876**	0.5805	0.5680	0.5467	0.5457	0.5425	0.5423
H_{11}	**0.6053**	0.5856	0.5728	0.5274	0.5188	0.4905	0.4913	0.4806	0.3600
H_{12}	0.6053	**0.6167**	0.5972	0.5967	0.5908	0.5839	0.5663	0.5610	0.5561
H_{13}	0.4356	**0.5780**	0.5366	0.4249	0.4289	0.4246	0.4391	0.4463	0.4349
H_{14}	**0.6053**	0.6000	0.4908	0.4525	0.4152	0.4107	0.3258	0.3805	0.3717
H_{15}	0.2629	0.2010	0.3974	**0.3886**	0.3786	0.3710	0.3868	0.3832	0.3862
AGNES - Manhattan Distance									
R_{21}	0.7763	0.7771	**0.7957**	0.7913	0.7886	0.7725	0.7842	0.7834	0.7809
R_{22}	0.7763	**0.7766**	0.7723	0.7725	0.7729	0.7726	0.7683	0.7687	0.7684
R_{23}	0.7082	**0.8021**	0.7968	0.7412	0.7457	0.7695	0.7672	0.7647	0.7579
R_{24}	0.7763	**0.8568**	0.8370	0.7815	0.7621	0.7858	0.7834	0.7671	0.7644
R_{25}	0.7763	**0.7992**	0.7562	0.7519	0.7486	0.7462	0.7782	0.7757	0.7907
J_{21}	**0.5951**	0.5866	0.5568	0.5473	0.5111	0.4738	0.4278	0.4256	0.4165
J_{22}	**0.5951**	0.5891	0.5811	0.5781	0.5752	0.5714	0.5633	0.5604	0.5495
J_{23}	0.4783	**0.5749**	0.5297	0.4011	0.3326	0.3455	0.3389	0.3321	0.3128
J_{24}	0.5951	**0.6549**	0.5719	0.4260	0.3734	0.3814	0.3744	0.3274	0.3163
J_{25}	**0.5951**	0.5723	0.4807	0.4298	0.4224	0.4168	0.4085	0.3992	0.3989
FM_{21}	**0.7715**	0.7600	0.7178	0.7096	0.6765	0.6433	0.6144	0.6126	0.6052
FM_{22}	**0.7715**	0.7635	0.7562	0.7523	0.7485	0.7440	0.7366	0.7329	0.7203
FM_{23}	0.6611	**0.7339**	0.6926	0.5747	0.5224	0.5572	0.5510	0.5445	0.5257
FM_{24}	0.7715	**0.7922**	0.7313	0.6113	0.5634	0.5956	0.5895	0.5466	0.5381
FM_{25}	**0.7715**	0.7324	0.6502	0.6022	0.5951	0.5897	0.5983	0.5908	0.6089
H_{21}	**0.6299**	0.6100	0.5625	0.5508	0.5195	0.4769	0.4831	0.4810	0.4741
H_{22}	**0.6299**	0.6159	0.6030	0.5967	0.5908	0.5839	0.5716	0.5663	0.5490
H_{23}	0.4417	**0.5841**	0.5408	0.3920	0.3770	0.4473	0.4408	0.4340	0.4146
H_{24}	0.6299	**0.6842**	0.6192	0.4764	0.4235	0.4949	0.4885	0.4441	0.4379
H_{25}	**0.6299**	0.5805	0.4648	0.4234	0.4146	0.4080	0.4669	0.4600	0.5066
DIANA - Euclidean Distance									
Rand	0.7637	0.8639	**0.8773**	0.8588	0.8032	0.7809	0.7953	0.7837	0.7703
Jaccard	0.5723	0.6624	**0.6852**	0.6076	0.4531	0.3894	0.3945	0.3584	0.3158
FM	0.7505	0.7970	**0.8132**	0.7631	0.6455	0.5914	0.6160	0.5864	0.5502
Hubert	0.5922	0.6949	**0.7219**	0.6716	0.5342	0.4757	0.5267	0.4976	0.4634
DIANA - Manhattan Distance									
Rand	0.7598	0.8511	**0.8690**	0.8475	0.8363	0.7861	0.8019	0.7885	0.7862
Jaccard	0.5653	0.6371	**0.6661**	0.6105	0.5418	0.4013	0.4124	0.3701	0.3634
FM	0.7440	0.7785	**0.7996**	0.7589	0.7165	0.6032	0.6316	0.5982	0.5925
Hubert	0.5809	0.6666	**0.7023**	0.6484	0.6176	0.4904	0.5440	0.5113	0.5056
PAM - Euclidean Distance									
Rand	0.7719	**0.8797**	0.8548	0.8360	0.8437	0.7878	0.7842	0.7656	0.7608
Jaccard	0.5872	**0.6959**	0.6024	0.5516	0.5465	0.3845	0.3626	0.3079	0.2930
FM	0.7642	**0.8208**	0.7577	0.7202	0.7269	0.6000	0.5881	0.5384	0.5246
Hubert	0.6167	**0.7305**	0.6615	0.6155	0.6400	0.5014	0.4971	0.4471	0.4338
PAM - Manhattan Distance									
Rand	0.7637	**0.8859**	0.8515	0.8480	0.8369	0.7817	0.7627	0.7698	0.7579
Jaccard	0.5723	**0.7084**	0.5959	0.5678	0.5322	0.3741	0.3195	0.3170	0.2784
FM	0.7505	**0.8294**	0.7524	0.7381	0.7140	0.5868	0.5362	0.5494	0.5148
Hubert	0.5922	**0.7439**	0.6536	0.6473	0.6218	0.4826	0.4301	0.4606	0.4290

Table E–1: External Criteria values, Iris Data Set

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES - Euclidean Distance									
R_{11}	0.5758	0.5768	**0.8889**	0.8888	0.8873	0.8863	0.8749	0.8704	0.8702
R_{12}	0.5453	0.5462	0.5470	0.5488	0.5507	0.5516	0.5525	0.5534	**0.5562**
R_{13}	0.8252	0.8067	0.8067	0.7851	0.8122	0.8112	**0.8581**	0.8538	0.8490
R_{14}	**0.9240**	0.9012	0.8659	0.8554	0.8550	0.6863	0.6827	0.6760	0.6749
R_{15}	0.5585	0.5585	0.5575	**0.7776**	0.7583	0.7582	0.7570	0.7552	0.7552
J_{11}	0.5491	0.5453	**0.8096**	0.8094	0.8057	0.8041	0.7841	0.7763	0.7760
J_{12}	0.5444	0.5444	0.5444	0.5445	0.5446	0.5446	**0.5448**	0.5447	0.5450
J_{13}	0.7364	0.7083	0.7055	0.6718	0.6956	0.6940	**0.7474**	0.7398	0.7310
J_{14}	**0.8677**	0.8264	0.7607	0.7417	0.7409	0.4345	0.4281	0.4122	0.4083
J_{15}	0.5416	0.5415	0.5400	**0.6708**	0.6408	0.6407	0.6389	0.6362	0.6362
FM_{11}	0.7328	0.7276	**0.8952**	0.8951	0.8931	0.8921	0.8802	0.8755	0.8754
FM_{12}	**0.7375**	0.7372	0.7369	0.7364	0.7358	0.7355	0.7353	0.7350	0.7343
FM_{13}	0.8494	0.8298	0.8276	0.8037	0.8212	0.8201	**0.8606**	0.8560	0.8508
FM_{14}	**0.9293**	0.9060	0.8687	0.8575	0.8571	0.6515	0.6465	0.6366	0.6348
FM_{15}	0.7291	0.7289	0.7273	**0.8035**	0.7812	0.7811	0.7797	0.7777	0.7777
H_{11}	0.1395	0.1364	**0.7790**	0.7788	0.7770	0.7753	0.7547	0.7467	0.7465
H_{12}	0.0213	0.0302	0.0371	0.0483	0.0575	0.0617	0.0657	0.0695	**0.0801**
H_{13}	0.6489	0.6097	0.6094	0.5664	0.6265	0.6247	**0.7360**	0.7289	0.7216
H_{14}	**0.8475**	0.8063	0.7499	0.7330	0.7327	0.4820	0.4767	0.4735	0.4747
H_{15}	0.0825	0.0821	0.0775	**0.5503**	0.5116	0.5115	0.5092	0.5056	0.5057
AGNES - Manhattan Distance									
R_{21}	0.5758	0.5810	**0.8986**	0.8985	0.8982	0.8871	0.8867	0.8840	0.8581
R_{22}	0.5453	0.5462	0.5479	0.5488	0.5507	0.5516	0.5525	0.5534	**0.5553**
R_{23}	**0.9267**	0.9018	0.8763	0.8781	0.8621	0.8610	0.8592	0.8502	0.8456
R_{24}	**0.9376**	0.9121	0.8828	0.8648	0.8633	0.8558	0.8538	0.8514	0.8503
R_{25}	0.5877	**0.9040**	0.9031	0.9025	0.9025	0.8741	0.8620	0.8613	0.8609
J_{21}	0.5491	0.5501	**0.8266**	0.8265	0.8257	0.8066	0.8059	0.8010	0.7555
J_{22}	0.5444	0.5444	0.5445	0.5445	0.5446	0.5446	0.5447	0.5447	**0.5449**
J_{23}	**0.8746**	0.8317	0.7874	0.7877	0.7590	0.7571	0.7540	0.7378	0.7292
J_{24}	**0.8909**	0.8459	0.7918	0.7593	0.7559	0.7421	0.7384	0.7342	0.7322
J_{25}	0.5527	**0.8326**	0.8310	0.8299	0.8299	0.7781	0.7565	0.7553	0.7544
FM_{21}	0.7328	0.7322	**0.9052**	0.9050	0.9047	0.8934	0.8930	0.8901	0.8628
FM_{22}	**0.7375**	0.7372	0.7366	0.7364	0.7358	0.7355	0.7353	0.7350	0.7345
FM_{23}	**0.9331**	0.9083	0.8820	0.8829	0.8659	0.8647	0.8628	0.8531	0.8480
FM_{24}	**0.9423**	0.9171	0.8866	0.8677	0.8660	0.8579	0.8557	0.8532	0.8520
FM_{25}	0.7333	**0.9092**	0.9083	0.9077	0.9077	0.8780	0.8652	0.8645	0.8639
H_{21}	0.1395	0.1519	**0.7972**	0.7971	0.7962	0.7756	0.7748	0.7701	0.7246
H_{22}	0.0213	0.0302	0.0430	0.0483	0.0575	0.0617	0.0657	**0.0695**	0.0767
H_{23}	**0.8520**	0.8034	0.7564	0.7629	0.7357	0.7339	0.7308	0.7163	0.7093
H_{24}	**0.8744**	0.8262	0.7770	0.7471	0.7459	0.7344	0.7310	0.7272	0.7255
H_{25}	0.1691	**0.8102**	0.8085	0.8073	0.8073	0.7595	0.7393	0.7381	0.7373
DIANA - Euclidean Distance									
Rand	**0.8896**	0.8482	0.8421	0.8311	0.8303	0.8298	0.8297	0.8266	0.8186
Jaccard	**0.8200**	0.7518	0.7418	0.7233	0.7219	0.7212	0.7210	0.7158	0.7024
FM	**0.9014**	0.8584	0.8520	0.8399	0.8390	0.8386	0.8385	0.8350	0.8262
Hubert	**0.7775**	0.6954	0.6839	0.6635	0.6620	0.6613	0.6611	0.6552	0.6411
DIANA - Manhattan Distance									
Rand	**0.8767**	0.8390	0.8337	0.8178	0.8519	0.8494	0.8473	0.8564	0.8558
Jaccard	**0.8021**	0.7408	0.7323	0.7058	0.7422	0.7377	0.7341	0.7458	0.7447
FM	**0.8906**	0.8511	0.8455	0.8279	0.8553	0.8526	0.8504	0.8590	0.8584
Hubert	**0.7518**	0.6760	0.6658	0.6361	0.7171	0.7129	0.7094	0.7309	0.7299
PAM - Euclidean Distance									
Rand	**0.9213**	0.8809	0.6966	0.6419	0.6208	0.6117	0.6145	0.6052	0.6018
Jaccard	**0.8662**	0.7967	0.4694	0.3606	0.3193	0.3024	0.2982	0.2844	0.2780
FM	**0.9283**	0.8873	0.6688	0.5867	0.5522	0.5369	0.5404	0.5247	0.5188
Hubert	**0.8411**	0.7634	0.4723	0.4062	0.3790	0.3657	0.3856	0.3657	0.3611
PAM - Manhattan Distance									
Rand	**0.9106**	0.8676	0.6804	0.6757	0.6674	0.6798	0.6180	0.6153	0.6128
Jaccard	**0.8498**	0.7766	0.4487	0.4281	0.4126	0.4154	0.3040	0.2989	0.2933
FM	**0.9189**	0.8747	0.6486	0.6386	0.6264	0.6417	0.5463	0.5418	0.5374
Hubert	**0.8196**	0.7366	0.4370	0.4470	0.4352	0.4851	0.3921	0.3882	0.3864

Table E–2: External Criteria values, Breastw Data Set

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES - Euclidean Distance									
R_{11}	0.5401	0.5417	0.5434	0.5451	0.5485	0.5520	0.5573	0.5573	**0.5591**
R_{12}	0.5401	0.5417	0.5434	0.5451	0.5468	0.5485	0.5502	0.5519	**0.5537**
R_{13}	0.5227	0.5039	0.5037	0.5099	0.5966	0.5993	**0.6003**	0.5997	0.5996
R_{14}	0.5989	0.6322	**0.6589**	0.6397	0.6195	0.5902	0.5865	0.5787	0.5806
R_{15}	0.5524	0.5574	0.5592	0.5580	0.5582	0.5585	**0.5622**	0.5616	0.5621
J_{11}	0.5384	0.5384	0.5383	0.5383	0.5383	0.5384	0.5388	0.5387	**0.5389**
J_{12}	0.5384	0.5384	0.5383	0.5383	0.5383	0.5383	0.5384	0.5384	**0.5385**
J_{13}	**0.4186**	0.3476	0.3359	0.3334	0.3405	0.3355	0.3356	0.3340	0.3339
J_{14}	0.4439	0.4517	**0.4660**	0.4103	0.3736	0.3132	0.3069	0.2920	0.2927
J_{15}	**0.4477**	0.4446	0.4447	0.4432	0.4431	0.4430	0.4432	0.4425	0.4424
FM_{11}	**0.7331**	0.7325	0.7318	0.7312	0.7300	0.7289	0.7274	0.7274	0.7269
FM_{12}	**0.7331**	0.7325	0.7318	0.7312	0.7306	0.7300	0.7294	0.7289	0.7283
FM_{13}	**0.5918**	0.5166	0.5044	0.5026	0.5349	0.5337	0.5344	0.5332	0.5330
FM_{14}	0.6152	0.6258	**0.6430**	0.6009	0.5685	0.5144	0.5080	0.4934	0.4953
FM_{15}	**0.6206**	0.6168	0.6167	0.6153	0.6151	0.6150	0.6150	0.6143	0.6142
H_{11}	0.0275	0.0392	0.0484	0.0563	0.0700	0.0821	0.0985	0.0984	**0.1035**
H_{12}	0.0275	0.0392	0.0484	0.0563	0.0634	0.0699	0.0761	0.0819	**0.0875**
H_{13}	0.0270	0.0100	0.0136	0.0291	0.2529	0.2659	**0.2689**	0.2686	0.2684
H_{14}	0.1979	0.2775	**0.3417**	0.3287	0.2971	0.2560	0.2493	0.2368	0.2428
H_{15}	0.0875	0.1004	0.1045	0.1022	0.1028	0.1034	0.1117	0.1107	**0.1121**
AGNES - Manhattan Distance									
R_{21}	0.5452	0.5503	0.5557	0.5556	0.5592	0.5592	0.5610	0.5647	**0.5744**
R_{22}	0.5401	0.5417	0.5434	0.5451	0.5468	0.5485	0.5502	0.5519	**0.5537**
R_{23}	0.5092	**0.5245**	0.5106	0.5033	0.5087	0.5069	0.5064	0.5097	0.5119
R_{24}	0.5727	**0.6381**	0.6265	0.6323	0.6175	0.5890	0.5934	0.5846	0.5805
R_{25}	0.5562	0.5911	0.5901	0.5897	**0.5939**	0.5933	0.5932	0.5932	0.5932
J_{21}	0.5384	0.5385	0.5388	0.5387	0.5390	0.5389	0.5391	0.5394	**0.5408**
J_{22}	0.5384	0.5384	0.5383	0.5383	0.5383	0.5383	0.5384	0.5384	**0.5385**
J_{23}	0.4059	**0.4109**	0.3427	0.3282	0.3152	0.3128	0.3120	0.3101	0.3110
J_{24}	0.4206	**0.4513**	0.4308	0.3891	0.3631	0.2976	0.2988	0.2824	0.2748
J_{25}	**0.5394**	0.4255	0.4196	0.4191	0.4208	0.4153	0.4153	0.4152	0.4151
FM_{21}	**0.7313**	0.7296	0.7280	0.7280	0.7270	0.7270	0.7266	0.7257	0.7242
FM_{22}	**0.7331**	0.7325	0.7318	0.7312	0.7306	0.7300	0.7294	0.7289	0.7283
FM_{23}	0.5790	**0.5833**	0.5120	0.4965	0.4842	0.4815	0.4806	0.4793	0.4806
FM_{24}	0.5924	**0.6268**	0.6088	0.5851	0.5614	0.5050	0.5096	0.4934	0.4857
FM_{25}	0.7286	**0.5981**	0.5928	0.5923	0.5943	0.5895	0.5894	0.5894	0.5893
H_{21}	0.0570	0.0769	0.0938	0.0936	0.1040	0.1038	0.1088	0.1185	**0.1422**
H_{22}	0.0275	0.0392	0.0484	0.0563	0.0634	0.0699	0.0761	0.0819	**0.0875**
H_{23}	−0.0007	0.0345	0.0274	0.0155	0.0326	0.0293	0.0283	0.0370	**0.0419**
H_{24}	0.1445	0.2935	0.2746	**0.3245**	0.3002	0.2671	0.2818	0.2673	0.2603
H_{25}	0.0964	0.1866	0.1867	0.1860	0.1955	**0.1966**	0.1965	0.1965	0.1964
DIANA - Euclidean Distance									
Rand	0.5401	0.5620	0.5618	0.5615	0.5613	0.5611	**0.5788**	0.5786	0.5639
Jaccard	0.5384	0.5401	0.5399	0.5396	0.5393	0.5392	**0.5419**	0.5417	0.3990
FM	**0.7331**	0.7275	0.7273	0.7270	0.7268	0.7266	0.7241	0.7239	0.5715
Hubert	0.0275	0.1128	0.1120	0.1108	0.1098	0.1091	**0.1529**	0.1521	0.1319
DIANA - Manhattan Distance									
Rand	0.5004	0.5092	0.5120	0.5323	0.5526	0.5519	**0.5558**	0.5557	0.5557
Jaccard	**0.4059**	0.4012	0.4017	0.3968	0.3958	0.3948	0.3961	0.3960	0.3959
FM	**0.5798**	0.5737	0.5741	0.5681	0.5676	0.5667	0.5681	0.5680	0.5680
Hubert	−0.0224	0.0016	0.0081	0.0581	0.1063	0.1050	**0.1138**	0.1136	0.1136
PAM - Euclidean Distance									
Rand	0.5865	**0.6249**	0.5459	0.5745	0.5648	0.5559	0.5497	0.5458	0.5308
Jaccard	0.4310	**0.4394**	0.2838	0.2893	0.2663	0.2395	0.2287	0.2209	0.1882
FM	0.6028	**0.6149**	0.4672	0.4885	0.4668	0.4427	0.4302	0.4216	0.3850
Hubert	0.1733	**0.2651**	0.1416	0.2251	0.2130	0.2068	0.1947	0.1878	0.1633
PAM - Manhattan Distance									
Rand	0.5452	**0.6063**	0.6047	0.5748	0.5639	0.5544	0.5471	0.5450	0.5542
Jaccard	0.3928	**0.4130**	0.3890	0.2938	0.2663	0.2365	0.2236	0.2187	0.2203
FM	0.5643	**0.5901**	0.5715	0.4915	0.4660	0.4395	0.4245	0.4194	0.4301
Hubert	0.0899	**0.2312**	0.2412	0.2223	0.2099	0.2044	0.1900	0.1872	0.2203

Table E–3: External Criteria values, Ionosphere Data Set

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES - Euclidean Distance									
R_{11}	0.3896	0.4416	0.4805	0.4762	0.6147	0.6364	0.6277	0.6234	**0.6667**
R_{12}	0.3333	0.4113	0.4545	0.4502	0.4719	0.5152	0.6147	0.6667	**0.6753**
R_{13}	0.4762	0.5628	0.5931	0.5931	**0.6970**	0.6797	0.6970	0.6926	0.6970
R_{14}	0.4762	0.5281	0.6277	0.6710	0.6710	0.6840	0.6667	0.7013	**0.7056**
R_{15}	0.3896	0.4372	0.4545	0.5152	0.6320	0.6364	0.6277	**0.6710**	0.6667
J_{11}	**0.2985**	0.2712	0.2683	0.2622	0.2458	0.2364	0.2037	0.1944	0.1250
J_{12}	0.2903	0.2804	0.2800	0.2743	0.2561	0.2583	0.2583	0.2870	**0.2925**
J_{13}	**0.3086**	0.2628	0.2598	0.2480	0.2473	0.2043	0.2135	0.1932	0.1954
J_{14}	**0.3086**	0.2876	0.2951	0.2083	0.2000	0.1978	0.1538	0.1481	0.1500
J_{15}	0.2985	**0.3011**	0.2800	0.2632	0.2342	0.2364	0.2037	0.1364	0.1250
FM_{11}	**0.5189**	0.4608	0.4477	0.4391	0.3950	0.3825	0.3395	0.3268	0.2441
FM_{12}	**0.5196**	0.4830	0.4719	0.4638	0.4305	0.4255	0.4117	0.4461	0.4527
FM_{13}	**0.5119**	0.4240	0.4158	0.3995	0.4053	0.3504	0.3684	0.3435	0.3485
FM_{14}	**0.5119**	0.4667	0.4587	0.3525	0.3424	0.3445	0.2828	0.2991	0.3058
FM_{15}	**0.5189**	0.5104	0.4719	0.4328	0.3797	0.3825	0.3395	0.2619	0.2441
H_{11}	0.0528	0.0207	0.0419	0.0285	**0.1132**	0.1251	0.0820	0.0674	0.0629
H_{12}	−0.0209	0.0190	0.0466	0.0326	0.0151	0.0497	0.1274	0.2077	**0.2221**
H_{13}	0.1183	0.0907	0.1106	0.0962	**0.2137**	0.1532	0.1902	0.1680	0.1781
H_{14}	0.1183	0.1044	**0.1810**	0.1430	0.1359	0.1554	0.0892	0.1583	0.1711
H_{15}	0.0528	0.0838	0.0466	0.0572	0.1183	**0.1251**	0.0820	0.0815	0.0629
AGNES - Manhattan Distance									
R_{21}	0.3896	0.4372	0.5584	0.6364	**0.6970**	0.6797	0.6753	0.6710	0.6883
R_{22}	0.3896	0.4372	0.4329	0.4502	0.4719	0.5152	0.6190	0.6494	**0.7186**
R_{23}	0.4762	0.5974	0.6147	0.6320	0.6147	0.7013	0.7186	**0.7229**	0.7186
R_{24}	0.4762	0.5281	0.6320	0.6667	0.7013	0.6840	0.6970	**0.7143**	0.7100
R_{25}	0.3896	0.5281	0.6147	0.6667	0.6797	**0.7143**	0.6970	0.6926	0.6883
J_{21}	0.2985	0.3011	0.2817	0.3000	**0.3137**	0.2745	0.2647	0.2549	0.1529
J_{22}	0.2985	**0.3011**	0.2957	0.2743	0.2561	0.2583	0.1776	0.1900	0.2262
J_{23}	0.3086	0.3008	**0.3101**	0.2609	0.2261	0.2069	0.1975	0.2000	0.1875
J_{24}	**0.3086**	0.2876	0.2342	0.2222	0.2247	0.1798	0.1765	0.1646	0.1519
J_{25}	0.2985	0.2876	**0.3101**	0.2222	0.2211	0.2235	0.1765	0.1647	0.1529
FM_{21}	**0.5189**	0.5104	0.4518	0.4640	0.4781	0.4320	0.4201	0.4080	0.2936
FM_{22}	**0.5189**	0.5104	0.5027	0.4638	0.4305	0.4255	0.3035	0.3244	0.3953
FM_{23}	**0.5119**	0.4711	0.4805	0.4140	0.3688	0.3637	0.3680	0.3750	0.3586
FM_{24}	**0.5119**	0.4667	0.3797	0.3682	0.3828	0.3232	0.3273	0.3313	0.3130
FM_{25}	**0.5189**	0.4667	0.4805	0.3682	0.3701	0.3895	0.3273	0.3107	0.2936
H_{21}	0.0528	0.0838	0.1142	0.1937	**0.2653**	0.2109	0.1970	0.1830	0.1303
H_{22}	0.0528	0.0838	0.0684	0.0326	0.0151	0.0497	0.0446	0.0957	**0.2423**
H_{23}	0.1183	0.1665	0.1903	0.1463	0.0916	0.1943	0.2292	**0.2421**	0.2251
H_{24}	0.1183	0.1044	0.1183	0.1487	**0.2057**	0.1417	0.1656	0.2032	0.1847
H_{25}	0.0528	0.1044	0.1903	0.1487	0.1665	**0.2312**	0.1656	0.1482	0.1303
DIANA - Euclidean Distance									
Rand	0.3333	0.3680	0.3983	0.4329	0.5195	0.5325	0.5498	0.6104	**0.6190**
Jaccard	**0.2903**	0.2808	0.2684	0.2599	0.2649	0.2394	0.2180	0.1509	0.1456
FM	**0.5196**	0.4943	0.4661	0.4445	0.4346	0.3947	0.3614	0.2652	0.2588
Hubert	−0.0209	−0.0142	−0.0176	−0.0072	**0.0629**	0.0355	0.0216	0.0055	0.0106
DIANA - Manhattan Distance									
Rand	0.3333	0.3680	0.4892	0.5065	0.5411	0.6104	0.6623	**0.7316**	0.7273
Jaccard	**0.2903**	0.2808	0.2484	0.2549	0.2148	0.2174	0.2121	0.2530	0.2410
FM	**0.5196**	0.4943	0.4153	0.4220	0.3575	0.3571	0.3550	0.4305	0.4161
Hubert	−0.0209	−0.0142	0.0155	0.0383	0.0099	0.0776	0.1338	**0.2844**	0.2692
PAM - Euclidean Distance									
Rand	0.3333	0.3680	0.3983	0.4762	0.5195	0.6364	0.6277	**0.6450**	0.6407
Jaccard	**0.2903**	0.2808	0.2684	0.2622	0.2649	0.2364	0.2110	0.1633	0.1531
FM	**0.5196**	0.4943	0.4661	0.4391	0.4346	0.3825	0.3491	0.2883	0.2734
Hubert	−0.0209	−0.0142	−0.0176	0.0285	0.0629	**0.1251**	0.0895	0.0640	0.0477
PAM - Manhattan Distance									
Rand	0.3333	0.3680	0.5801	0.6494	0.6840	0.6753	0.6710	0.6926	**0.7013**
Jaccard	**0.2903**	0.2808	0.2707	0.2636	0.2772	0.2424	0.2165	0.2283	0.2333
FM	**0.5196**	0.4943	0.4324	0.4173	0.4357	0.3940	0.3623	0.3828	0.3920
Hubert	−0.0209	−0.0142	0.1142	0.1665	**0.2186**	0.1779	0.1499	0.1929	0.2114

Table E–4: External Criteria values, Breastcc Data Set

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES - Euclidean Distance									
R_{11}	0.2842	0.2960	0.3052	0.3241	0.3333	0.3441	0.3953	0.3938	**0.4071**
R_{12}	0.2842	0.2960	0.3052	0.3169	0.3292	0.3379	0.3635	0.3769	**0.3907**
R_{13}	0.2954	0.2960	0.5397	0.5899	0.6293	0.6293	0.6682	0.6687	**0.6994**
R_{14}	0.4378	0.6272	0.6682	0.7143	0.7143	0.7389	0.7394	0.7430	**0.7721**
R_{15}	0.2842	0.2960	0.5325	0.5376	0.5433	0.5540	0.6175	0.6257	**0.6278**
J_{11}	**0.2692**	0.2651	0.2593	0.2601	0.2543	0.2496	0.2386	0.2366	0.2341
J_{12}	**0.2692**	0.2651	0.2593	0.2552	0.2514	0.2452	0.2388	0.2360	0.2337
J_{13}	0.2650	**0.2651**	0.2483	0.2201	0.1735	0.1669	0.1828	0.1768	0.1859
J_{14}	0.2257	0.2401	0.2448	**0.2687**	0.2530	0.2619	0.2580	0.2607	0.2645
J_{15}	**0.2692**	0.2651	0.2230	0.2182	0.2141	0.2153	0.2162	0.2199	0.2106
FM_{11}	**0.5111**	0.5003	0.4872	0.4841	0.4713	0.4602	0.4292	0.4262	0.4190
FM_{12}	**0.5111**	0.5003	0.4872	0.4768	0.4670	0.4537	0.4366	0.4288	0.4218
FM_{13}	0.5002	**0.5003**	0.4145	0.3645	0.2960	0.2865	0.3125	0.3045	0.3243
FM_{14}	0.3985	0.3890	0.3933	0.4261	0.4083	0.4267	0.4230	0.4277	**0.4573**
FM_{15}	**0.5111**	0.5003	0.3769	0.3687	0.3616	0.3619	0.3564	0.3611	0.3482
H_{11}	−0.0324	−0.0339	−0.0460	**−0.0225**	−0.0334	−0.0381	−0.0266	−0.0321	−0.0281
H_{12}	**−0.0324**	−0.0339	−0.0460	−0.0468	−0.0462	−0.0559	−0.0500	−0.0460	−0.0407
H_{13}	−0.0353	−0.0339	0.0783	0.0693	0.0449	0.0370	0.0986	0.0929	**0.1446**
H_{14}	−0.0241	0.1241	0.1649	0.2392	0.2255	0.2704	0.2687	0.2770	**0.3475**
H_{15}	−0.0324	−0.0339	0.0334	0.0295	0.0272	0.0366	0.0861	**0.0976**	0.0881
AGNES - Manhattan Distance									
R_{21}	0.2842	0.2960	0.3154	0.3241	0.3333	0.4403	0.4414	0.5878	**0.5975**
R_{22}	0.2842	0.2960	0.3082	0.3169	0.3292	0.3472	0.3728	0.3799	**0.3968**
R_{23}	0.3077	0.3088	0.5791	0.6057	0.6390	0.6400	0.6411	**0.6810**	0.6790
R_{24}	0.5607	0.5607	0.6764	0.7225	0.7230	0.7471	**0.7650**	0.7609	0.7619
R_{25}	0.2842	0.4685	0.4752	0.5837	0.5873	0.5914	0.6037	0.6078	**0.6083**
J_{21}	**0.2692**	0.2651	0.2662	0.2601	0.2543	0.2097	0.2077	0.1901	0.1889
J_{22}	**0.2692**	0.2651	0.2613	0.2552	0.2514	0.2518	0.2457	0.2384	0.2385
J_{23}	0.2616	0.2619	**0.2641**	0.2488	0.1850	0.1720	0.1665	0.1649	0.1573
J_{24}	0.2753	0.2114	0.2565	**0.2821**	0.2699	0.2605	0.2585	0.2357	0.2365
J_{25}	**0.2692**	0.2236	0.2170	0.2352	0.2331	0.2334	0.2110	0.2127	0.2130
FM_{21}	**0.5111**	0.5003	0.4975	0.4841	0.4713	0.3713	0.3678	0.3209	0.3184
FM_{22}	**0.5111**	0.5003	0.4903	0.4768	0.4670	0.4634	0.4467	0.4322	0.4288
FM_{23}	0.4909	0.4912	**0.4304**	0.4037	0.3126	0.2944	0.2867	0.2913	0.2805
FM_{24}	**0.4510**	0.3552	0.4082	0.4427	0.4294	0.4299	0.4438	0.4209	0.4228
FM_{25}	**0.5111**	0.3892	0.3771	0.3873	0.3838	0.3837	0.3501	0.3523	0.3525
H_{21}	−0.0324	−0.0339	−0.0066	−0.0225	−0.0334	−0.0520	−0.0548	0.0277	**0.0343**
H_{22}	−0.0324	−0.0339	−0.0342	−0.0468	−0.0462	−0.0292	−0.0258	−0.0382	**−0.0256**
H_{23}	−0.0337	−0.0314	**0.1255**	0.1198	0.0684	0.0543	0.0490	0.0968	0.0860
H_{24}	0.1341	0.0357	0.1856	0.2613	0.2517	0.2847	**0.3253**	0.3061	0.3094
H_{25}	−0.0324	−0.0074	−0.0146	0.0858	0.0854	**0.0887**	0.0681	0.0737	0.0744
DIANA - Euclidean Distance									
Rand	0.4849	0.6006	0.5883	0.5842	0.5853	0.5853	0.6318	0.6503	**0.6544**
Jaccard	**0.2840**	0.2353	0.2024	0.1944	0.1892	0.1843	0.1640	0.1589	0.1604
FM	**0.4837**	0.3852	0.3388	0.3275	0.3197	0.3124	0.2825	0.2770	0.2797
Hubert	**0.1216**	0.0979	0.0444	0.0305	0.0243	0.0178	0.0362	0.0507	0.0574
DIANA - Manhattan Distance									
Rand	0.4695	0.4941	0.5003	0.6221	0.6590	0.7409	0.7389	0.7435	**0.7558**
Jaccard	0.2228	0.2165	0.2161	0.1845	0.2005	**0.2481**	0.2284	0.2316	0.2404
FM	0.3877	0.3731	0.3714	0.3116	0.3349	0.4140	0.3929	0.3995	**0.4189**
Hubert	−0.0082	−0.0026	0.0008	0.0512	0.1070	0.2651	0.2485	0.2603	**0.2940**
PAM - Euclidean Distance									
Rand	0.4798	0.6318	0.7025	0.7609	0.7691	0.8126	**0.8331**	0.8249	0.8274
Jaccard	0.2463	0.2580	0.2949	0.3423	0.3309	0.3786	**0.4062**	0.3770	0.3805
FM	0.4238	0.4131	0.4555	0.5141	0.5072	0.5803	**0.6277**	0.6036	0.6106
Hubert	0.0412	0.1503	0.2509	0.3602	0.3668	0.4859	**0.5583**	0.5335	0.5439
PAM - Manhattan Distance									
Rand	0.5105	0.6406	0.6959	0.7783	0.7773	0.8018	**0.8075**	0.7926	0.8049
Jaccard	0.2431	0.2524	0.2852	**0.3679**	0.3338	0.3507	0.3461	0.3005	0.3245
FM	0.4122	0.4046	0.4438	0.5437	0.5147	0.5509	**0.5578**	0.5109	0.5449
Hubert	0.0528	0.1498	0.2345	0.4036	0.3843	0.4510	**0.4700**	0.4191	0.4637

Table E–5: External Criteria values, Srbct Data Set

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES - Euclidean Distance									
R_{11}	0.2369	0.2625	0.2915	0.3217	0.3531	0.3856	0.4053	0.4402	**0.4890**
R_{12}	0.2230	0.2532	0.2846	0.3217	0.3531	0.3856	0.4216	0.4425	**0.4739**
R_{13}	0.5947	0.6295	0.6388	0.6585	0.6667	0.8409	0.8502	**0.8595**	0.8513
R_{14}	0.6504	0.7526	0.8316	0.8351	0.8560	0.8839	0.8827	**0.8850**	0.8849
R_{15}	0.2369	0.2625	0.2915	0.3217	0.3531	0.3856	0.5006	0.5168	**0.5610**
J_{11}	0.2017	0.1982	0.1974	0.1978	0.1997	0.2033	0.1975	**0.2046**	0.2043
J_{12}	0.1930	0.1922	0.1927	0.1978	0.1997	0.2033	0.2108	0.2066	**0.2122**
J_{13}	0.2834	0.2974	0.2883	0.2933	0.2948	0.4606	0.4713	**0.4851**	0.4506
J_{14}	0.3266	0.4067	0.4948	0.5000	0.4609	**0.5146**	0.5073	0.5050	0.5025
J_{15}	0.2017	0.1982	0.1974	0.1978	0.1997	0.2033	**0.2389**	0.2325	0.2333
FM_{11}	**0.4459**	0.4324	0.4239	0.4178	0.4143	0.4137	0.3973	0.4027	0.3899
FM_{12}	**0.4298**	0.4214	0.4154	0.4178	0.4143	0.4137	0.4194	0.4060	0.4084
FM_{13}	0.4971	0.5067	0.4883	0.4885	0.4879	0.6332	0.6418	**0.6536**	0.6213
FM_{14}	0.5454	0.6127	0.6774	0.6814	0.6310	**0.6819**	0.6761	0.6762	0.6747
FM_{15}	**0.4459**	0.4324	0.4239	0.4178	0.4143	0.4137	0.4507	0.4339	0.4222
H_{11}	0.0693	0.0391	0.0385	0.0454	0.0578	0.0749	0.0583	0.0852	**0.0914**
H_{12}	−0.0131	−0.0010	0.0131	0.0454	0.0578	0.0749	0.1026	0.0914	**0.1103**
H_{13}	0.2841	0.3075	0.2856	0.2929	0.2951	0.5338	0.5479	**0.5657**	0.5288
H_{14}	0.3672	0.4799	0.5794	0.5851	0.5415	**0.6123**	0.6064	0.6095	0.6083
H_{15}	0.0693	0.0391	0.0385	0.0454	0.0578	0.0749	**0.1851**	0.1662	0.1687
AGNES - Manhattan Distance									
R_{21}	0.2230	0.2857	0.3182	0.4332	0.4634	0.4623	0.4797	0.7027	**0.7282**
R_{22}	0.2230	0.2532	0.2846	0.3171	0.3508	0.3856	0.4216	0.4541	**0.4739**
R_{23}	0.5308	0.6539	0.7131	0.7166	0.7143	0.8467	0.8641	0.8630	**0.8722**
R_{24}	0.5308	0.6980	0.7991	0.8235	0.8444	0.8479	0.8688	0.8688	**0.8780**
R_{25}	0.2230	0.3972	0.4042	0.4797	0.4797	0.4785	0.4797	0.7073	**0.7201**
J_{21}	0.1930	0.1940	0.1959	0.2229	0.2274	0.2258	0.2195	0.2948	**0.2931**
J_{22}	0.1930	0.1922	0.1927	0.1945	0.1980	0.2033	0.2108	**0.2167**	0.2122
J_{23}	0.2747	0.3348	0.3584	0.3613	0.3560	0.4677	0.4777	0.4732	**0.4884**
J_{24}	0.2477	0.3103	0.3663	0.3896	0.4071	0.4126	0.4488	0.4461	**0.4615**
J_{25}	0.1930	0.2184	0.2204	0.2222	0.2209	0.2191	0.2195	0.2800	**0.2891**
FM_{21}	0.4298	0.4179	0.4146	0.4400	0.4401	0.4373	0.4206	**0.4744**	0.4636
FM_{22}	**0.4298**	0.4214	0.4154	0.4120	0.4113	0.4137	0.4194	0.4222	0.4084
FM_{23}	0.5060	0.5571	0.5642	0.5666	0.5600	0.6390	0.6467	0.6426	**0.6573**
FM_{24}	0.4571	0.4992	0.5386	0.5608	0.5798	0.5857	0.6255	0.6236	**0.6428**
FM_{25}	0.4298	0.4409	0.4430	0.4257	0.4232	0.4202	0.4206	0.4513	**0.4605**
H_{21}	−0.0131	0.0200	0.0365	0.1425	0.1545	0.1495	0.1310	0.2939	**0.2943**
H_{22}	−0.0131	−0.0010	0.0131	0.0303	0.0508	0.0749	0.1026	**0.1220**	0.1103
H_{23}	0.2833	0.3848	0.4072	0.4112	0.4020	0.5430	0.5627	0.5580	**0.5795**
H_{24}	0.2055	0.3210	0.4123	0.4505	0.4851	0.4936	0.5497	0.5483	**0.5758**
H_{25}	−0.0131	0.1297	0.1364	0.1386	0.1348	0.1299	0.1310	0.2701	**0.2867**
DIANA - Euclidean Distance									
Rand	0.5679	0.6144	0.6249	0.8060	0.8130	0.8095	0.8235	0.8223	**0.8304**
Jaccard	0.2634	0.2860	0.2774	0.3993	0.4015	0.3903	0.4039	0.4000	**0.4065**
FM	0.4724	0.4939	0.4757	0.5757	0.5760	0.5641	0.5763	0.5722	**0.5782**
Hubert	0.2404	0.2854	0.2643	0.4547	0.4589	0.4446	0.4656	0.4608	**0.4722**
DIANA - Manhattan Distance									
Rand	0.5679	0.5784	0.6098	0.6585	0.6655	0.8049	0.8130	0.8130	**0.8223**
Jaccard	0.2634	0.2546	0.2648	0.2650	0.2653	0.3538	0.3611	0.3586	**0.3704**
FM	0.4724	0.4537	0.4602	0.4442	0.4424	0.5231	0.5307	0.5279	**0.5406**
Hubert	0.2404	0.2179	0.2384	0.2388	0.2399	0.4008	0.4140	0.4113	**0.4305**
PAM - Euclidean Distance									
Rand	0.5761	0.7596	0.8653	0.8873	**0.8943**	0.8850	0.8862	0.8885	0.8850
Jaccard	0.2829	0.3689	0.5486	0.5611	**0.5767**	0.5374	0.5333	0.5340	0.5194
FM	0.5035	0.5562	0.7167	0.7190	**0.7316**	0.6993	0.6965	0.6979	0.6859
Hubert	0.2888	0.4122	0.6357	0.6486	**0.6658**	0.6284	0.6270	0.6306	0.6169
PAM - Manhattan Distance									
Rand	0.6051	0.7712	0.8397	0.8502	0.8641	0.8792	0.8955	0.8966	**0.9071**
Jaccard	0.2872	0.3863	0.4889	0.4580	0.4823	0.5185	0.5610	0.5572	**0.5812**
FM	0.4996	0.5741	0.6652	0.6285	0.6508	0.6832	0.7199	0.7179	**0.7404**
Hubert	0.2906	0.4375	0.5678	0.5348	0.5664	0.6088	0.6564	0.6561	**0.6869**

Table E–6: External Criteria, Brain Data Set

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES - Euclidean Distance									
R_{11}	0.5321	0.5321	0.5243	0.5168	0.5638	0.5642	0.5642	**0.5959**	0.5951
R_{12}	0.5321	0.5243	0.5325	0.5246	0.5329	0.5415	0.5505	**0.5700**	0.5606
R_{13}	**0.6197**	0.6088	0.5383	0.5689	0.6033	0.6033	0.5657	0.5622	0.5610
R_{14}	**0.6999**	0.6060	0.6686	0.6185	0.5829	0.5794	0.5548	0.5509	0.5509
R_{15}	0.5321	**0.6228**	0.6227	0.6119	0.5450	0.5391	0.5391	0.5395	0.5395
J_{11}	**0.5275**	0.5142	0.5012	0.4884	0.4930	0.4932	0.4649	0.3659	0.3603
J_{12}	**0.5275**	0.5148	0.5144	0.5014	0.5010	0.5011	0.5015	0.5043	0.4891
J_{13}	**0.4737**	0.4562	0.3292	0.3235	0.3153	0.3097	0.2276	0.2213	0.2159
J_{14}	**0.5504**	0.4027	0.4254	0.3313	0.2623	0.2561	0.2025	0.1927	0.1870
J_{15}	**0.5275**	0.5218	0.5190	0.5015	0.3761	0.3646	0.3337	0.3339	0.3316
FM_{11}	**0.7206**	0.7032	0.6888	0.6746	0.6688	0.6690	0.6376	0.5520	0.5479
FM_{12}	**0.7206**	0.7064	0.7034	0.6890	0.6862	0.6839	0.6821	0.6806	0.6650
FM_{13}	**0.6430**	0.6269	0.5036	0.5094	0.5278	0.5254	0.4496	0.4425	0.4388
FM_{14}	**0.7107**	0.5825	0.6288	0.5492	0.4846	0.4779	0.4249	0.4154	0.4136
FM_{15}	**0.7206**	0.6892	0.6863	0.6699	0.5482	0.5366	0.5079	0.5082	0.5061
H_{11}	−0.0365	−0.0060	−0.0236	−0.0371	0.1014	0.1023	0.1092	0.2337	**0.2354**
H_{12}	−0.0365	−0.0510	−0.0044	−0.0222	0.0101	0.0386	0.0649	**0.1152**	0.0944
H_{13}	0.2365	0.2171	0.1005	0.1849	0.3035	**0.3102**	0.2641	0.2575	0.2599
H_{14}	0.4021	0.2377	**0.4153**	0.3424	0.2896	0.2833	0.2512	0.2479	0.2570
H_{15}	−0.0365	0.2329	**0.2330**	0.2108	0.0965	0.0871	0.1004	0.1013	0.1023
AGNES - Manhattan Distance									
R_{21}	0.5321	0.5407	0.5497	**0.5591**	0.5501	0.4984	0.4879	0.4879	0.4894
R_{22}	0.5321	0.5407	0.5325	0.5411	0.5501	**0.5696**	0.5603	0.5513	0.5606
R_{23}	**0.5931**	0.5829	0.5360	0.5344	0.5614	0.5317	0.5618	0.5274	0.5227
R_{24}	**0.6197**	0.5493	0.5798	0.5423	0.5599	0.5544	0.5352	0.5352	0.5317
R_{25}	0.5321	0.4937	0.4781	0.4812	0.4832	0.4797	0.4832	0.4832	**0.5599**
J_{21}	0.5275	0.5276	0.5281	**0.5290**	0.5150	0.3399	0.3100	0.3034	0.3040
J_{22}	0.5275	**0.5276**	0.5144	0.5145	0.5150	0.5178	0.5029	0.4882	0.4891
J_{23}	**0.4424**	0.4260	0.3444	0.3378	0.3456	0.2602	0.2463	0.1871	0.1655
J_{24}	**0.4737**	0.3510	0.3471	0.2600	0.2363	0.2257	0.1846	0.1796	0.1629
J_{25}	**0.5275**	0.3749	0.3461	0.3406	0.3408	0.3363	0.3311	0.2641	0.2485
FM_{21}	**0.7206**	0.7179	0.7156	0.7138	0.6988	0.5083	0.4762	0.4693	0.4702
FM_{22}	**0.7206**	0.7179	0.7034	0.7008	0.6988	0.6967	0.6813	0.6660	0.6650
FM_{23}	**0.6136**	0.5980	0.5171	0.5107	0.5241	0.4416	0.4538	0.3823	0.3623
FM_{24}	**0.6430**	0.5257	0.5317	0.4474	0.4464	0.4349	0.3897	0.3869	0.3734
FM_{25}	**0.7206**	0.5456	0.5142	0.5083	0.5086	0.5036	0.4982	0.4280	0.4531
H_{21}	−0.0365	0.0213	0.0585	**0.0888**	0.0607	0.0005	−0.0135	−0.0112	−0.0078
H_{22}	−0.0365	0.0213	−0.0044	0.0308	0.0607	**0.1143**	0.0909	0.0708	0.0944
H_{23}	0.1849	0.1671	0.0880	0.0870	0.1516	0.1176	**0.2258**	0.1563	0.1594
H_{24}	**0.2365**	0.1179	0.2006	0.1489	0.2293	0.2199	0.1898	0.1958	0.2020
H_{25}	−0.0365	−0.0247	−0.0496	−0.0400	−0.0355	−0.0420	−0.0319	−0.0095	**0.2164**
DIANA - Euclidean Distance									
Rand	**0.9190**	0.9022	0.8549	0.6764	0.6655	0.6483	0.6154	0.5696	0.5665
Jaccard	**0.8596**	0.8303	0.7443	0.4225	0.4029	0.3713	0.3126	0.2237	0.2136
FM	**0.9245**	0.9075	0.8570	0.6356	0.6196	0.5933	0.5401	0.4540	0.4466
Hubert	**0.8373**	0.8050	0.7238	0.4503	0.4334	0.4079	0.3553	0.2873	0.2889
DIANA - Manhattan Distance									
Rand	0.5321	0.7050	**0.7496**	0.7222	0.6283	0.6311	0.5763	0.5728	0.5728
Jaccard	0.5275	0.5567	**0.5868**	0.5416	0.3498	0.3510	0.2495	0.2432	0.2304
FM	0.7206	0.7159	**0.7453**	0.7115	0.5656	0.5689	0.4716	0.4648	0.4606
Hubert	−0.0365	0.4121	**0.5200**	0.4738	0.3569	0.3661	0.2791	0.2727	0.2912
PAM - Euclidean Distance									
Rand	**0.8689**	0.8294	0.6811	0.6268	0.6092	0.5861	0.5724	0.5579	0.5540
Jaccard	**0.7776**	0.6953	0.4402	0.3380	0.2900	0.2507	0.2131	0.1894	0.1816
FM	**0.8753**	0.8283	0.6446	0.5603	0.5270	0.4855	0.4568	0.4268	0.4182
Hubert	**0.7392**	0.6873	0.4443	0.3665	0.3628	0.3204	0.3195	0.2888	0.2828
PAM - Manhattan Distance									
Rand	**0.7782**	0.6510	0.6577	0.6146	0.5966	0.5716	0.5595	0.5485	0.5446
Jaccard	**0.6469**	0.4358	0.4120	0.3160	0.2790	0.2289	0.1986	0.1786	0.1715
FM	**0.7864**	0.6218	0.6147	0.5400	0.5073	0.4584	0.4313	0.4073	0.3984
Hubert	**0.5591**	0.3440	0.3909	0.3472	0.3241	0.2883	0.2800	0.2595	0.2519

Table E–7: External Criteria values, Leukemia Data Set

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES - Euclidean Distance									
R_{11}	0.8842	**0.8942**	0.7467	0.7425	0.7530	0.7493	0.6330	0.6267	0.6166
R_{12}	**0.4923**	0.4818	0.4717	0.8535	0.8340	0.8149	0.7790	0.7784	0.7890
R_{13}	**0.6727**	0.5373	0.6393	0.6679	0.6388	0.6346	0.6013	0.5854	0.5891
R_{14}	0.8842	**0.9482**	0.7626	0.7044	0.6452	0.6198	0.6166	0.5955	0.5923
R_{15}	0.8842	**0.8942**	0.7086	0.7192	0.7129	0.6325	0.6219	0.5754	0.5717
J_{11}	0.7993	**0.8133**	0.5528	0.5453	0.5557	0.5490	0.3397	0.3283	0.3102
J_{12}	0.4869	0.4706	0.4544	**0.7414**	0.7068	0.6732	0.6097	0.6088	0.6204
J_{13}	**0.5052**	0.3006	0.3467	0.3598	0.3038	0.2956	0.2314	0.2008	0.2023
J_{14}	0.7993	**0.8990**	0.5371	0.4237	0.3082	0.2588	0.2456	0.2040	0.1977
J_{15}	0.7993	**0.8133**	0.4855	0.4948	0.4833	0.3387	0.3197	0.2360	0.2293
FM_{11}	0.8888	**0.8972**	0.7196	0.7139	0.7242	0.7193	0.5418	0.5307	0.5125
FM_{12}	0.6902	0.6710	0.6519	**0.8517**	0.8289	0.8063	0.7617	0.7610	0.7711
FM_{13}	**0.6713**	0.4690	0.5506	0.5854	0.5341	0.5262	0.4598	0.4247	0.4314
FM_{14}	0.8888	**0.9473**	0.7273	0.6431	0.5443	0.4960	0.4887	0.4437	0.4365
FM_{15}	0.8888	**0.8972**	0.6667	0.6774	0.6683	0.5409	0.5221	0.4315	0.4236
H_{11}	0.7698	**0.7890**	0.5111	0.5035	0.5288	0.5223	0.3151	0.3031	0.2839
H_{12}	−0.0616	−0.0868	−0.1059	**0.7077**	0.6702	0.6346	0.5688	0.5679	0.5921
H_{13}	0.3455	0.0799	0.3304	0.4233	**0.3730**	0.3655	0.3033	0.2713	0.2867
H_{14}	0.7698	**0.8983**	0.5849	0.4925	0.3947	0.3499	0.3548	0.3157	0.3096
H_{15}	0.7698	**0.7890**	0.4434	0.4697	0.4586	0.3141	0.2940	0.2003	0.1923
AGNES - Manhattan Distance									
R_{21}	0.8842	**0.8942**	0.7356	0.7462	0.6319	0.6208	0.6023	0.6066	0.5976
R_{22}	0.4923	**0.8942**	0.8736	0.8535	0.8340	0.8149	0.7964	0.8070	0.7890
R_{23}	**0.9154**	0.7631	0.6282	0.6060	0.5775	0.5912	0.6319	0.6224	0.5970
R_{24}	0.8842	**0.9482**	0.8006	0.6737	0.6420	0.6134	0.6103	0.6007	0.5896
R_{25}	0.8842	**0.8942**	0.7255	0.7361	0.6261	0.6071	0.5944	0.5796	0.5717
J_{21}	0.7993	**0.8133**	0.5331	0.5433	0.3378	0.3178	0.2845	0.2770	0.2604
J_{22}	0.4869	**0.8133**	0.7768	0.7414	0.7068	0.6732	0.6405	0.6527	0.6204
J_{23}	**0.8506**	0.5817	0.3436	0.3044	0.2540	0.2502	0.2704	0.2516	0.2013
J_{24}	0.7993	**0.8990**	0.6113	0.3639	0.3021	0.2464	0.2331	0.2144	0.1925
J_{25}	0.7993	**0.8133**	0.5154	0.5252	0.3273	0.2931	0.2702	0.2436	0.2293
FM_{21}	0.8888	**0.8972**	0.7045	0.7150	0.5400	0.5202	0.4857	0.4855	0.4676
FM_{22}	0.6902	**0.8972**	0.8744	0.8517	0.8289	0.8063	0.7836	0.7935	0.7711
FM_{23}	**0.9200**	0.7413	0.5405	0.5009	0.4458	0.4555	0.5186	0.5000	0.4468
FM_{24}	0.8888	**0.9473**	0.7775	0.5940	0.5385	0.4832	0.4756	0.4553	0.4305
FM_{25}	0.8888	**0.8972**	0.6907	0.7012	0.5297	0.4948	0.4703	0.4404	0.4236
H_{21}	0.7698	**0.7890**	0.4913	0.5168	0.3131	0.2920	0.2559	0.2762	0.2583
H_{22}	−0.0616	**0.7890**	0.7472	0.7077	0.6702	0.6346	0.6005	0.6242	0.5921
H_{23}	**0.8335**	0.5403	0.2971	0.2542	0.1963	0.2442	0.3904	0.3741	0.3288
H_{24}	0.7698	**0.8983**	0.6455	0.4427	0.3893	0.3383	0.3433	0.3257	0.3044
H_{25}	0.7698	**0.7890**	0.4735	0.4992	0.3021	0.2653	0.2400	0.2093	0.1923
DIANA - Euclidean Distance									
Rand	**0.8842**	0.6917	0.6155	0.5859	0.5685	0.5785	0.5637	0.5743	0.6267
Jaccard	**0.7993**	0.4656	0.3336	0.2823	0.2521	0.2558	0.2297	0.2341	0.2584
FM	**0.8888**	0.6473	0.5260	0.4721	0.4380	0.4480	0.4173	0.4292	0.5083
Hubert	**0.7698**	0.4064	0.2651	0.2065	0.1703	0.1985	0.1668	0.1980	0.3841
DIANA - Manhattan Distance									
Rand	**0.8842**	0.7065	0.6113	0.6214	0.6044	0.6150	0.5722	0.6245	0.6203
Jaccard	**0.7993**	0.4913	0.3263	0.3315	0.3016	0.3073	0.2303	0.2542	0.2458
FM	**0.8888**	0.6686	0.5186	0.5285	0.4979	0.5096	0.4247	0.5042	0.4958
Hubert	**0.7698**	0.4331	0.2570	0.2840	0.2511	0.2808	0.1935	0.3805	0.3732
PAM - Euclidean Distance									
Rand	0.8842	**0.9217**	0.7240	0.6716	0.6436	0.6245	0.6097	0.6029	0.5965
Jaccard	0.7993	**0.8491**	0.4679	0.3598	0.2986	0.2612	0.2320	0.2185	0.2044
FM	0.8888	**0.9195**	0.6734	0.5904	0.5407	0.5045	0.4745	0.4599	0.4459
Hubert	0.7698	**0.8474**	0.5156	0.4392	0.4016	0.3689	0.3424	0.3297	0.3205
PAM - Manhattan Distance									
Rand	0.8842	**0.9217**	0.7240	0.6764	0.6372	0.6103	0.6087	0.6039	0.5976
Jaccard	0.7993	**0.8491**	0.4679	0.3704	0.2942	0.2418	0.2300	0.2190	0.2065
FM	0.8888	**0.9195**	0.6734	0.5985	0.5299	0.4770	0.4723	0.4620	0.4482
Hubert	0.7698	**0.8474**	0.5156	0.4453	0.3789	0.3302	0.3404	0.3343	0.3225

Table E–8: External Criteria values, Lymphoma Data Set

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES - Euclidean Distance									
R_{11}	0.4958	0.5094	**0.5110**	0.5096	0.5084	0.5075	0.5075	0.5075	0.5057
R_{12}	0.4958	0.4954	0.4960	0.4968	0.4978	0.4989	0.5003	0.5018	**0.5083**
R_{13}	0.4958	**0.5094**	0.5063	0.5079	0.5065	0.5046	0.5038	0.5038	0.5036
R_{14}	0.5020	0.5067	0.5053	0.5112	**0.5127**	0.5104	0.5100	0.5098	0.5088
R_{15}	0.4958	0.5094	**0.5110**	0.5096	0.5084	0.5075	0.5075	0.5075	0.5075
J_{11}	**0.4907**	0.4282	0.4285	0.4266	0.4248	0.4233	0.4214	0.4208	0.4136
J_{12}	**0.4907**	0.4853	0.4807	0.4762	0.4717	0.4674	0.4632	0.4591	0.4377
J_{13}	**0.4907**	0.4374	0.2994	0.2997	0.2974	0.2939	0.2924	0.2912	0.2909
J_{14}	**0.3364**	0.3038	0.3013	0.2124	0.2124	0.2077	0.1803	0.1799	0.1762
J_{15}	**0.4907**	0.4374	0.4380	0.4362	0.4346	0.4332	0.3158	0.3145	0.3136
FM_{11}	**0.6971**	0.6109	0.6111	0.6089	0.6070	0.6053	0.6030	0.6022	0.5937
FM_{12}	**0.6971**	0.6897	0.6831	0.6766	0.6703	0.6641	0.6581	0.6522	0.6231
FM_{13}	**0.6971**	0.6225	0.4624	0.4629	0.4602	0.4561	0.4544	0.4531	0.4528
FM_{14}	**0.5034**	0.4673	0.4644	0.3693	0.3699	0.3640	0.3339	0.3335	0.3288
FM_{15}	**0.6971**	0.6225	0.6228	0.6208	0.6191	0.6176	0.4805	0.4791	0.4781
H_{11}	0.0029	0.0261	**0.0295**	0.0263	0.0236	0.0213	0.0211	0.0210	0.0165
H_{12}	0.0029	−0.0009	0.0010	0.0031	0.0055	0.0081	0.0111	0.0144	**0.0252**
H_{13}	0.0029	0.0277	0.0113	**0.0144**	0.0116	0.0076	0.0060	0.0059	0.0055
H_{14}	0.0042	0.0122	0.0094	0.0203	**0.0238**	0.0185	0.0178	0.0173	0.0149
H_{15}	0.0029	0.0277	**0.0312**	0.0279	0.0251	0.0228	0.0141	0.0141	0.0140
AGNES - Manhattan Distance									
R_{21}	0.4958	**0.5094**	0.5063	0.5079	0.5065	0.5053	0.5051	0.5042	0.5042
R_{22}	0.4958	0.4954	0.4960	0.4968	0.4978	0.4970	**0.5057**	0.5042	0.5034
R_{23}	0.5046	0.5036	0.5048	0.5061	0.5061	0.5079	0.5137	**0.5139**	0.5127
R_{24}	0.5046	0.5059	0.5098	0.5086	0.5145	**0.5158**	0.5117	0.5117	0.5117
R_{25}	0.4958	0.5036	0.5048	0.5048	0.5061	0.5065	**0.5067**	0.5055	0.5053
J_{21}	**0.4907**	0.4374	0.3052	0.3055	0.3032	0.3012	0.2984	0.2966	0.2941
J_{22}	**0.4907**	0.4853	0.4807	0.4762	0.4717	0.4659	0.4262	0.4193	0.4181
J_{23}	**0.3401**	0.3357	0.3024	0.2990	0.2984	0.2018	0.1896	0.1886	0.1866
J_{24}	**0.3401**	0.3048	0.2132	0.2103	0.1984	0.1934	0.1700	0.1583	0.1575
J_{25}	**0.4907**	0.3393	0.3371	0.3266	0.3079	0.3003	0.3004	0.2987	0.2912
FM_{21}	**0.6971**	0.6225	0.4687	0.4692	0.4665	0.4642	0.4612	0.4591	0.4563
FM_{22}	**0.6971**	0.6897	0.6831	0.6766	0.6703	0.6626	0.6092	0.6009	0.5996
FM_{23}	**0.5076**	0.5026	0.4656	0.4619	0.4613	0.3567	0.3458	0.3448	0.3421
FM_{24}	**0.5076**	0.4683	0.3697	0.3662	0.3556	0.3508	0.3234	0.3104	0.3095
FM_{25}	**0.6971**	0.5067	0.5042	0.4924	0.4718	0.4634	0.4635	0.4616	0.4533
H_{21}	0.0029	**0.0277**	0.0114	0.0145	0.0118	0.0094	0.0089	0.0069	0.0068
H_{22}	0.0029	−0.0009	0.0010	0.0031	0.0055	0.0021	**0.0180**	0.0138	0.0121
H_{23}	0.0093	0.0072	0.0083	0.0108	0.0108	0.0126	**0.0266**	0.0271	0.0243
H_{24}	0.0093	0.0107	0.0171	0.0145	0.0283	**0.0317**	0.0224	0.0227	0.0228
H_{25}	0.0029	0.0074	0.0096	0.0091	0.0111	0.0117	**0.0121**	0.0097	0.0090
DIANA - Euclidean Distance									
Rand	0.5145	0.5129	0.5116	0.5057	0.5164	0.5176	**0.5199**	0.5180	0.5191
Jaccard	**0.3814**	0.3779	0.3613	0.2680	0.1920	0.1911	0.1886	0.1811	0.1814
FM	**0.5543**	0.5503	0.5314	0.4280	0.3496	0.3492	0.3477	0.3387	0.3397
Hubert	0.0311	0.0278	0.0241	0.0092	0.0332	0.0362	**0.0422**	0.0378	0.0407
DIANA - Manhattan Distance									
Rand	**0.5145**	0.5129	0.5116	0.5057	0.5110	0.5110	0.5114	0.5129	0.5116
Jaccard	**0.3749**	0.3712	0.3520	0.2680	0.1835	0.1779	0.1780	0.1779	0.1751
FM	**0.5468**	0.5426	0.5210	0.4280	0.3379	0.3317	0.3321	0.3327	0.3290
Hubert	**0.0306**	0.0273	0.0236	0.0092	0.0201	0.0202	0.0212	0.0251	0.0217
PAM - Euclidean Distance									
Rand	0.5046	0.5216	0.5279	0.5556	**0.5562**	0.5446	0.5401	0.5387	0.5455
Jaccard	**0.3475**	0.2901	0.2326	0.2230	0.2062	0.1840	0.1581	0.1547	0.1403
FM	**0.5160**	0.4542	0.3964	0.4011	0.3871	0.3586	0.3311	0.3267	0.3213
Hubert	0.0097	0.0423	0.0582	0.1310	**0.1386**	0.1113	0.1064	0.1032	0.1349
PAM - Manhattan Distance									
Rand	0.5075	0.5230	0.5152	0.5442	0.5434	0.5420	**0.5583**	0.5546	0.5517
Jaccard	**0.3412**	0.2814	0.2185	0.2165	0.1850	0.1818	0.1760	0.1603	0.1467
FM	**0.5088**	0.4452	0.3772	0.3883	0.3587	0.3546	0.3647	0.3486	0.3346
Hubert	0.0151	0.0452	0.0295	0.1013	0.1075	0.1044	**0.1605**	0.1570	0.1552

Table E–9: External Criteria values, Prostate Data Set

Index	c = 2	c = 3	c = 4	c = 5	c = 6	c = 7	c = 8	c = 9	c = 10
AGNES - Euclidean Distance									
R_{11}	**0.5177**	0.5172	0.5172	0.5093	0.4796	0.4892	0.4902	0.4902	0.4955
R_{12}	0.5256	0.5172	0.5262	0.5262	0.5357	0.5457	0.5362	0.5463	**0.5568**
R_{13}	0.5003	0.5003	0.4844	0.4759	0.4759	0.4749	0.4881	**0.5013**	0.5013
R_{14}	0.5003	0.5479	**0.5605**	0.5262	0.5145	0.5145	0.5124	0.5114	0.5034
R_{15}	0.4923	0.4981	0.4929	0.4929	0.4939	0.4939	0.4950	0.4913	**0.5045**
J_{11}	**0.4908**	0.4899	0.4897	0.4586	0.3668	0.2764	0.2768	0.2714	0.2667
J_{12}	**0.5201**	0.5057	0.5047	0.4886	0.4880	0.4881	0.4717	0.4717	0.4723
J_{13}	**0.3868**	0.2680	0.2060	0.1917	0.1714	0.1627	0.1388	0.1380	0.1341
J_{14}	**0.3762**	0.3214	0.3098	0.2133	0.1862	0.1715	0.1557	0.1429	0.1289
J_{15}	**0.4402**	0.3303	0.2222	0.2159	0.2051	0.1944	0.1948	0.1889	0.1664
FM_{11}	**0.6787**	0.6778	0.6774	0.6406	0.5373	0.4408	0.4415	0.4359	0.4324
FM_{12}	**0.7146**	0.6982	0.6942	0.6737	0.6706	0.6682	0.6508	0.6487	0.6475
FM_{13}	**0.5586**	0.4349	0.3670	0.3484	0.3265	0.3164	0.2991	0.3105	0.3068
FM_{14}	**0.5470**	0.4981	0.4931	0.3968	0.3652	0.3527	0.3375	0.3258	0.3046
FM_{15}	**0.6216**	0.4979	0.3869	0.3806	0.3703	0.3597	0.3606	0.3527	0.3396
H_{11}	−0.0257	−0.0269	−0.0265	−0.0261	−0.0540	−0.0033	−0.0010	0.0006	**0.0144**
H_{12}	−0.0383	−0.0533	−0.0064	0.0064	0.0358	0.0634	0.0430	0.0686	**0.0939**
H_{13}	−0.0130	0.0255	0.0042	−0.0139	−0.0092	−0.0100	0.0370	0.0834	**0.0861**
H_{14}	−0.0090	0.1234	**0.1630**	0.1192	0.0986	0.1080	0.1119	0.1183	0.0982
H_{15}	−0.0632	0.0015	0.0213	0.0232	0.0295	0.0332	**0.0360**	0.0277	0.0772
AGNES - Manhattan Distance									
R_{21}	**0.5177**	0.5172	0.5093	0.5093	0.4796	0.4892	0.4902	0.4865	0.4865
R_{22}	0.5256	0.5352	0.5352	0.5452	0.5357	0.5457	0.5362	**0.5463**	0.5368
R_{23}	**0.5003**	0.4939	0.4754	0.4754	0.4765	0.4892	0.4812	0.4828	0.4775
R_{24}	0.4939	0.4781	0.4876	0.5019	**0.5061**	0.5061	0.5061	0.4955	0.4960
R_{25}	**0.5003**	0.4939	0.4833	0.4833	0.4828	0.4886	0.4950	0.4960	0.4960
J_{21}	**0.4908**	0.4899	0.4589	0.4586	0.3668	0.2764	0.2768	0.2716	0.2661
J_{22}	**0.5201**	0.5197	0.5040	0.5040	0.4880	0.4881	0.4717	0.4717	0.4549
J_{23}	**0.3868**	0.2610	0.2006	0.1935	0.1674	0.1534	0.1402	0.1268	0.1147
J_{24}	**0.3708**	0.2640	0.2471	0.2010	0.1969	0.1593	0.1509	0.1327	0.1273
J_{25}	**0.4694**	0.4390	0.2981	0.2976	0.2856	0.2641	0.2574	0.2578	0.1699
FM_{21}	**0.6787**	0.6778	0.6410	0.6406	0.5373	0.4408	0.4415	0.4353	0.4296
FM_{22}	**0.7146**	0.7109	0.6906	0.6878	0.6706	0.6682	0.6508	0.6487	0.6311
FM_{23}	**0.5586**	0.4262	0.3578	0.3501	0.3223	0.3151	0.2955	0.2819	0.2637
FM_{24}	**0.5413**	0.4257	0.4104	0.3704	0.3690	0.3345	0.3269	0.2995	0.2947
FM_{25}	**0.6561**	0.6199	0.4626	0.4621	0.4493	0.4281	0.4229	0.4236	0.3366
H_{21}	−0.0257	−0.0269	−0.0263	−0.0261	−0.0540	−0.0033	**−0.0010**	−0.0081	−0.0067
H_{22}	−0.0383	0.0203	0.0279	0.0591	0.0358	0.0634	0.0430	**0.0686**	0.0499
H_{23}	−0.0130	0.0123	−0.0173	−0.0157	−0.0068	**0.0344**	0.0146	0.0244	0.0111
H_{24}	−0.0218	−0.0258	0.0010	0.0529	0.0669	0.0867	**0.0921**	0.0654	0.0707
H_{25}	−0.0671	−0.0574	−0.0227	−0.0226	−0.0205	−0.0011	0.0159	0.0184	**0.0488**
DIANA - Euclidean Distance									
Rand	**0.5256**	0.4876	0.5050	0.4955	0.5040	0.5061	0.5124	0.5124	0.5066
Jaccard	**0.5047**	0.3308	0.2530	0.2135	0.2058	0.2038	0.1819	0.1812	0.1549
FM	**0.6943**	0.4978	0.4217	0.3794	0.3761	0.3754	0.3599	0.3592	0.3310
Hubert	−0.0086	−0.0228	0.0421	0.0309	0.0570	0.0638	0.0943	**0.0947**	0.0913
DIANA - Manhattan Distance									
Rand	**0.5558**	0.5558	0.5457	0.4992	0.5024	0.5024	0.5029	0.5029	0.4950
Jaccard	**0.3961**	0.3493	0.2702	0.1954	0.1846	0.1789	0.1725	0.1703	0.1571
FM	**0.5679**	0.5246	0.4546	0.3635	0.3550	0.3497	0.3441	0.3420	0.3232
Hubert	0.1129	0.1310	**0.1439**	0.0476	0.0615	0.0641	0.0689	0.0700	0.0511
PAM - Euclidean Distance									
Rand	**0.7250**	0.6219	0.5447	0.5204	0.5177	0.5061	0.5198	0.5108	0.5182
Jaccard	**0.5972**	0.4012	0.2603	0.2174	0.1835	0.1593	0.1569	0.1371	0.1398
FM	**0.7480**	0.5855	0.4463	0.3963	0.3656	0.3345	0.3467	0.3206	0.3325
Hubert	**0.4461**	0.2780	0.1467	0.0990	0.1108	0.0867	0.1393	0.1213	0.1503
PAM - Manhattan Distance									
Rand	0.4923	0.5108	**0.5405**	0.5246	0.5156	0.5082	0.5093	0.5050	0.5045
Jaccard	**0.3384**	0.2802	0.2566	0.2093	0.1843	0.1606	0.1439	0.1317	0.1267
FM	**0.5061**	0.4497	0.4409	0.3923	0.3645	0.3376	0.3242	0.3089	0.3040
Hubert	−0.0143	0.0472	**0.1364**	0.1167	0.1032	0.0932	0.1091	0.1023	0.1044

Table E–10: External Criteria values, Colon Data Set

REFERENCE LIST

[1] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. ISBN 3-900051-07-0.

[2] C.L. Blake D.J. Newman, S. Hettich and C.J. Merz. UCI repository of machine learning databases, 1998.

[3] N. Bolshakova and F. Azuaje. Clustering genomic expression data: Design and evaluation principles. *Understanding and Using Micriarray Analysis Techniques: A Practical Guide*, 2002.

[4] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, 1999.

[5] H. Romesburg. *Cluster analysis for researchers*. Lifetime Learning Publications, 1984.

[6] C.D. Michener R.R. Sokal. A statistical method for evaluating systematic relationships. *U. Kansas Sci. Bull.*, 38:1409–1438, 1958.

[7] L. Kaufman and P. Rousseeuw. *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons, 1990.

[8] A. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1998.

[9] M. Murty A. Jain and P. Flyn. Data clustering: A review. *ACM Computing Surveys*, 31(3), 1999.

[10] J. Hartigan. *Clustering Algorithms*. John Wiley & Sons, 1975.

[11] J. Hartigan and M. Wong. A k-means clustering algorithm. *Journal of Applied Statistics*, 28, 1979.

[12] A. Afifi and V. Clark. *Computer-Aided Multivariate Analysis*. Champan and Hall, 1990.

[13] Y. Batistakis M. Halkidi and M. Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information System*, 17:2/3, 2001.

[14] Y. Batistakis M. Halkidi and M. Vazirgiannis. Cluster validity methods: Part i. *Sigmod Record*, 31(12), 2002.

[15] E. Fowlkes and C. Mallows. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Asociation*, 78, 1983.

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

[17] D. Davies and D. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 1979.

[18] J. Dunn. Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4, 1974.

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

[20] F. Azuaje. A cluster validity framework for genome expression data. *Bioinformatics*, 18(2), 2002.

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

[22] G. Milligan. An examination of of the efect of six types of error perturbation on fifteen clustering algorithms. *Psycometrika*, 45(3), 1980.

[23] G. Milligan. A monte carlo study of thirty internal criterion measures for cluster analysis. *Psycometrika*, 46(2), 1981.

[24] G. Milligan and M. Cooper. An examination of procedures for determining the number of cluster in a data set. *Psycometrika*, 50, 1985.