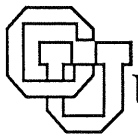
Clustering Objects On-line

Athman Bouguettaya

CU-CS-590-92 April 1992



University of Colorado at Boulder
DEPARTMENT OF COMPUTER SCIENCE

Clustering Objects On-line[1]

Athman Bouguettaya

CU-CS-590-92 April 1992

Department of Computer Science

Campus Box 430

University of Colorado at Boulder

Boulder, Colorado 80309-0430 USA

[1]This technical report is based on research conducted as part of the author's MS thesis.

Abstract

In this paper, we present the results of experiments conducted on three widely used clustering methods. In this analysis we focus on the stability of each method and the behavior of each one vis-a-vis the others. The input data are taken randomly from the segment [0,1] using few distributions. Various correlation coefficients are computed to help us understand each method's behavior. Surprisingly, the results came out to indicate that there is almost no difference among the chosen methods with respect to the input space. Furthermore, all methods appear to be almost stable. Those results show that the behavior is practically the same regardless of the method used.

1. Introduction and Motivation

Clustering has many applications that spans many areas of science and technology. This term is often used to define how two sets of entities are similar or dissimilar. The types of entities span a wide range that would include animals, plants, planets, and database objects.

Clustering methods differ in the *approach* used to measure the distance between two clusters (sets) of entities. In essence, every methods is different from other methods in the way entities are clustered.

Depending on the area of application, analyzing clusters is important in determining similar characteristics and behavior of clusters. We are particularly interested in database clustering as a test study. In databases, the problem is to minimize the number of disk accesses while fetching an object. In object-oriented databases, the problem is more specific. The existence of semantic relationship enables database designers to make some educated guess about future queries. This is due to the fact objects have some semantic relationships like *Is-Part-Of* and *IS-A* relationships [HuK88] [TsN91].

The idea is to adaptively cluster database objects based on previous query patterns. In the past, databases have relied on an off-line clustering [HuK88]. This is done after bringing the database down and make it unavailable for a certain period of time. In many instances and application, this is undesirable and counter-productive. Therefore the idea is to cluster database objects *while* it is servicing requests and queries. We would like to cluster objects in an *adaptive* fashion and *on-line*.

Conducting a cluster analysis on objects enables designers to have an insight about how objects should be clustered based on the requests made previously. This has the important feature that clustering is done *dynamically and adaptively* and without any human intervention. Most importantly, the database re-clusters objects on-line. The work presented in this paper is based on research conducted in [Bou87].

1.1. Clustering Description

In this study, we describe three of the most widely known and used *Cluster Analysis* methods. The difference among those methods resides mainly in the way how two clusters are joined. There are several other important methods which are also used in this area. The interested reader may find a more exhaustive study in [Rom84], [SnS73], and [SnS63].

1.2. Background

We start by giving some definitions of the terminology used in this paper.

Cluster Definition

A **cluster** is an ordered list of points. The points belong to an interval [a,b].[2] A **cluster** can also be defined as a set with the constraint that its elements are ordered.

Example of Clusters:

{0.1}, {0.6} are two different clusters

{0.2, 0.3, 0.5} is a cluster

Distance Between Two Clusters

The distance between two clusters involves some or all elements of the two clusters. The clustering method determines how the distance should be computed.

Coefficient of Correlation and Standard Deviation

The coefficient of correlation r of two random variables X and Y where

$$X = (x_1, x_2, x_3, \ldots, x_n) \tag{1.1}$$

and

$$Y = (y_1, y_2, y_3, \ldots, y_n) \tag{1.2}$$

is given by the following formula:

$$r = \frac{E(X,Y) - E(X)\,E(Y)}{(E(X^2) - E^2(X))^{1/2}\ (E(Y^2) - E^2(Y))^{1/2}} \tag{1.3}$$

where

$$E(X) = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1.4}$$

and

$$E(Y) = \frac{\sum_{i=1}^{n} y_i}{n} \tag{1.5}$$

[2] In our study the interval [a,b] is the interval [0,1]

and

$$E(X,Y) = \frac{\sum_{i=1}^{n} x_i\, y_i}{n} \tag{1.6}$$

All other terms are computed in a similar way.

The standard deviation of a random variable X is given by the following formula:

$$\sigma(X) = \sqrt{E(X^2) - E^2(X)} \tag{1.7}$$

Dissimilarity Coefficient

The dissimilarity coefficient of two clusters is the distance between those two clusters. The *smaller* the value of a *dis*similarity coefficient, the *more similar* two clusters are. The larger its value, the more dissimilar they are.[3]

Description of Some Clustering Methods

Cluster Analysis is the name of the various mathematical methods that are used in Numerical Classification to find the similarity among objects in a given set. The process of clustering is aimed at grouping OTU's (Operational Taxonomic Unit) progressively according to their similarity.

1) SLINK clustering method: SLINK is short for "Single LINKage" clustering method. When this method is used, two clusters are joined based upon the criterion that their dissimilarity coefficient is that of their nearest pair of elements, each one exactly in one cluster. This method is also called "nearest neighbor" clustering method.

2) CLINK clustering method: CLINK stands for "Complete LINKage" clustering method. Using this method we join two clusters when their dissimilarity coefficient is that of their furthest pair of elements, each one exactly in one cluster. This method is also called "furthest neighbor" clustering method.

3) UPGMA clustering method: UPGMA stands for "Unweighted Pair-Group Method using Arithmetic averages" clustering method. This method forms clusters based on the average value of dissimilarity coefficient between two clusters. This method is also called "average linkage" clustering method.[4]

---

[3] We will use interchangeably the terms "distance between clusters" and "dissimilarity coefficient" throughout the remaining paper.

[4] We will use "average method" to indicate this method for the remaining paper.

1.3. Statement of the Problem

In this section we will state *what* it is to be done. We are interested in finding the relationship among some clustering methods, namely **Slink, Clink** and **Average** using some parameters as means for the comparison.

Working Set and Distribution Functions

The working set of the three clustering methods is the set of real numbers taken on line in the interval [0..1].[5] The data are generated randomly using some distribution. Two kinds of distribution have been selected to carry out our random number generation [Knu71].

1- Uniform distribution of random numbers. The respective distribution function is the following:

$$F(x) = x \tag{1.8}$$

The schematic representation of this distribution function is:

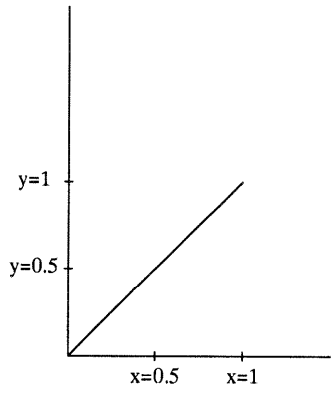


Fig 1.1 Illustration of the uniform distribution function

The density function of this distribution is:

$$f(x) = F'(x) = 1 \quad \text{for all } x \text{ such that } 0 \leq x \leq 1 \tag{1.9}$$

It is schematically represented as follows:

[5]From now on we will be using as working set the domain [0..1] unless explicitly stated otherwise.

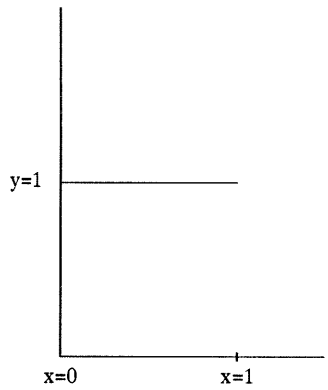


Fig 1.2 Illustration of the density of the uniform distribution

2- the second kind of distribution is given by the following function:

$$F(x) \equiv \begin{cases} 0.05 & \text{if } 0 \leq x < 0.37 \\ 0.475 & \text{if } 0.37 \leq x < 0.62 \\ 0.525 & \text{if } 0.62 \leq x < 0.743 \\ 0.95 & \text{if } 0.743 \leq x < 0.89 \\ 1 & \text{if } 0.89 < x \leq 1 \end{cases} \tag{1.10}$$

The schematic representation of this distribution is:

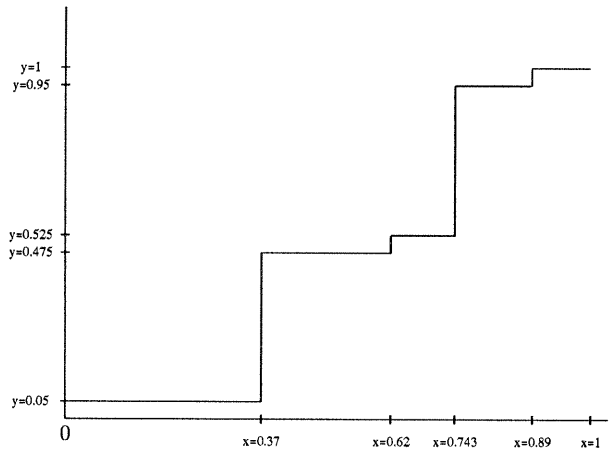


Fig 1.3 Illustration of the distribution function (1.10)

The density function of this discontinuous distribution can be obtained in similar way for all intervals as shown below. For an interval [a,b] the density function is:

$$f(x) = \frac{F(b) - F(a)}{b - a} \qquad \text{for all } x \text{ such that } a \leq x < b \qquad \textbf{(1.11)}$$

The density function of this distribution is schematically represented as follows:

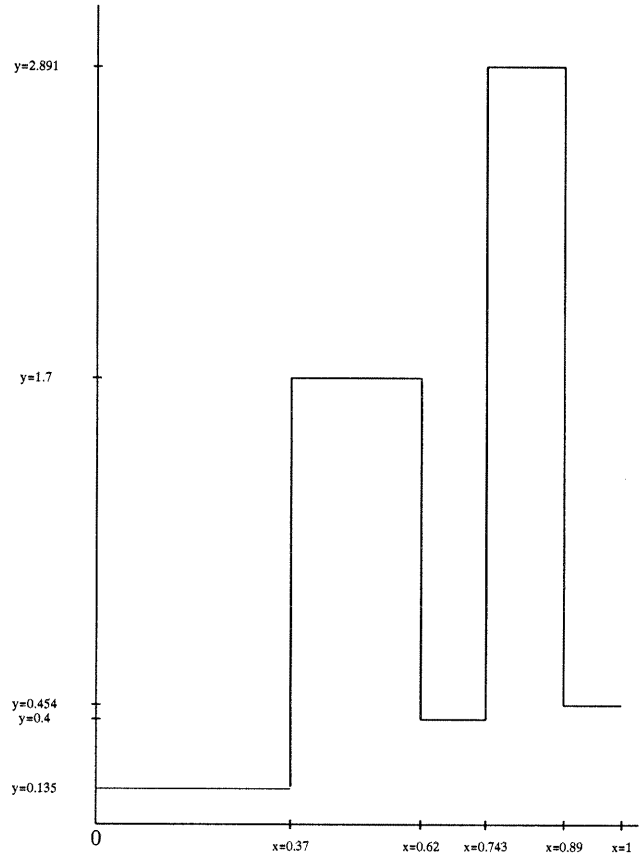


Fig 1.4 Illustration of the density of the distribution function (1.10)

Clustering and Tree Construction

After generating the random numbers we proceed by building the tree. As a step towards this goal the tree, we need to sort the data so that we can make use of the distance difference. The way of joining clusters varies accordingly to the method we are using. The distance between two clusters is computed using the dissimilarity coefficient.

General algorithm for clustering: Initially every cluster is composed of exactly one datum point. This algorithm is applicable to Slink, Clink or Average.

step 1:

Scan all the clusters and look for the minimum dissimilarity coefficient

step 2:

Scan all the clusters and look for the dissimilarity coefficient which is equal to the minimum and then join those clusters.

step 3:

If exactly one cluster remains then end

else go to step 1

We see here that the three methods differ only in the way of computing the dissimilarity coefficient.

There is a case where an ambiguity arises when clustering using Clink or Average method. Suppose when performing step 1 that we find three (3) successive clusters to be joined. When performing step 2 we would first join the two (2) first clusters. However when computing the dissimilarity coefficient between this new cluster and the former third cluster we would obtain a dissimilarity coefficient different than the minimum. In this case a question arises: How should we proceed? Two answers can be drawn:

(1) Either proceed by joining clusters using a recomputation of the dissimilarity coefficient each time we are in step 2.

(1) Or join all those clusters that have the dissimilarity coefficient equal to the minimum at once. In this case, we do not recompute the dissimilarity coefficient in step 2.

The issue of choosing one solution over another is irrelevant since there is no satisfactory distinction between those two solutions.

Following are examples of how we cluster some data. In example 1 we use Slink method. In example 2 we use Clink method. Finally in example 3 we use Average method. We assign an identification number to each point generated. We then proceed by sorting those points in an increasing order. The same sample is used as input for the three clustering methods. The dissimilarity coefficient is the magnitude of the difference between two points. The number of points is equal to 10.

The sorted data along with their identifications are:

Tab 1.1 Sample data used for clustering

Value	Id
0.1058909	4
0.2117760	8
0.3294196	3
0.4353047	7
0.5529483	2
0.6588334	6
0.7647185	10
0.7764770	1
0.8823621	5
0.9882473	9

example 1 (Slink):

step 1:

The clusters (10) and (1) are joined at the distance 0.0117584

step 2:

The clusters (6) and (10 1) are joined at the distance 0.1058850

step 3:

The clusters (4) and (8) are joined at the distance 0.105885140

step 4:

At the distance 0.105885148 the clusters (3) and (7) are merged to form one cluster as well as the clusters (2), (6 10 1), (5) and (6) are joined to form one cluster

step 5:

The clusters (4 8), (3 7) and (2 6 10 1 5 9) join to form one cluster at the distance 0.117643.

From the previous clustering we derive the following tree:

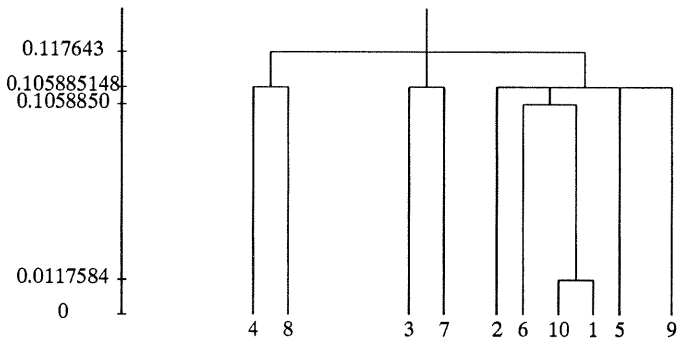


Fig 1.5 First example of a tree using Slink

example 2 (Clink):

step 1:

The clusters (10) and (1) are joined at the distance 0.0117584

step 2:

The clusters (4) and (8) are joined at the distance 0.105885140

step 3:

At the distance 0.105885148 the clusters (3) and (7) are merged to form one cluster as well as the clusters (2), (6) are merged to form one cluster. Same case for the clusters (5) and (9)

step 4:

The clusters (2 6) and (10 1) are joined at the distance 0.2235286

step 5:

The clusters (4 8), (3 7) join to form one cluster at the distance 0.3294138

step 6:

The clusters (2 6 10 1), (5 9) join to form one cluster at the distance 0.4352989

step 7:

The clusters (4 8 3 7) and (2 6 10 1 5 9) join to form one cluster at the distance 0.8823563.

From the previous clustering we derive the following tree:

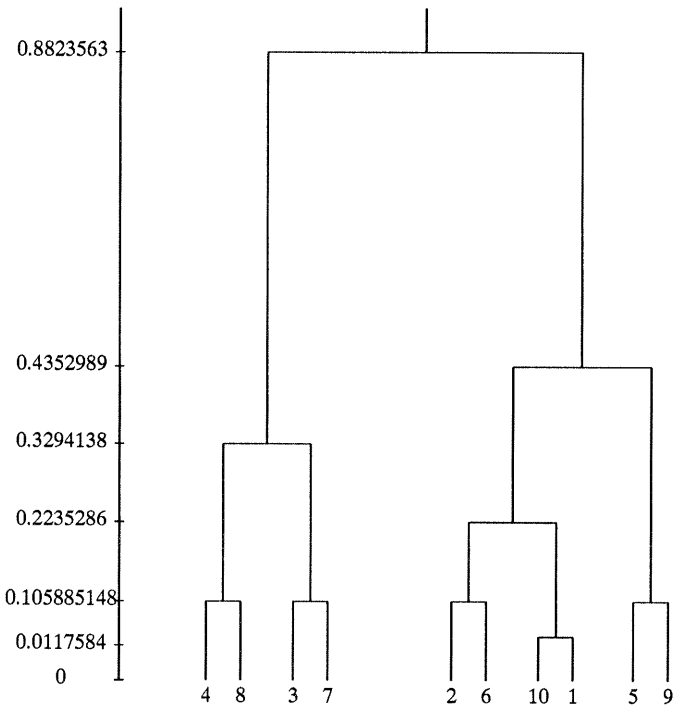


Fig 1.6 Second example of a tree using Clink

example 3 (Average):

step 1:

The clusters (10) and (1) are joined at the distance 0.0117584

step 2:

The clusters (4) and (8) are joined at the distance 0.105885140

step 3

At the distance 0.105885148 the clusters (3) and (7) are merged to form one cluster as well as the clusters (2), (6) are merged to form one cluster. Same case for the clusters (5) and (9)

step 4:

The clusters (2 6) and (10 1) are joined at the distance 0.1647068

step 5:

0.2235286

step 6:

The clusters (2 6 10 1), (5 9) join to form one cluster at the distance 0.2470603

step 7:

The clusters (4 8 3 7) and (2 6 10 1 5 9) join to form one cluster at the distance 0.49999992. From the previous clustering we derive the following tree:

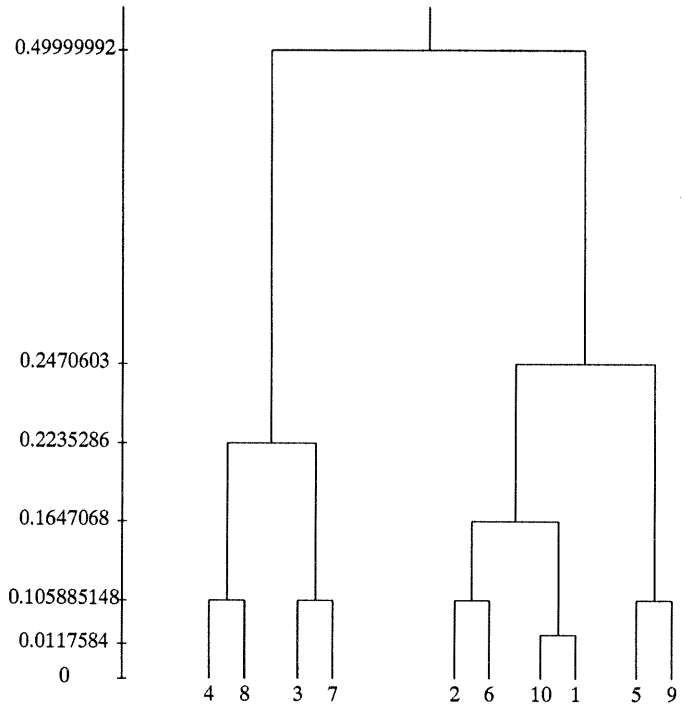


Fig 1.7 Third example of a tree using Average

Correlation Coefficients

The correlation coefficient is always computed between two lists. Let us show how the pairs of a list are prepared to be later used to compute the correlation coefficient [Knu71]. The first list consists of a sequence of distances between pairs of elements. The second list is also a sequence of distances between pairs of elements. The intersection of elements involved in computing distances in the first list and those involved in computing distances in the second list might be empty or non-empty.

example: As an example we want to compute the first correlation coefficient (see below for definition).

Tab 1.2 First data sample

Value	Id
0.7764770	1
0.5529483	2
0.3294196	3
0.1058909	4
0.8823621	5
0.6588334	6
0.4353047	7
0.2117760	8
0.9882473	9
0.7647185	10

Tab 1.3 Second data sample

Value	Id
0.7764770	1
0.5529483	2
0.3294196	3
0.1058909	4
0.8823621	5

after we sort the first sample we get the following list:

Tab 1.4 First sample data after sorting

Value	Id
0.1058909	4
0.2117760	8
0.3294196	3
0.4353047	7
0.5529483	2
0.6588334	6
0.7647185	10
0.7764770	1
0.8823621	5
0.9882473	9

after we sort the second sample we get the following list:

Tab 1.5 Second sample data after sorting

Value	Id
0.1058909	4
0.3294196	3
0.5529483	2
0.7764770	1
0.8823621	5

After having sorted the two samples we construct the trees of the two (2) lists. The tree of first list is shown in Fig 1.5 (using Slink method). The second tree is constructed in a similar way.

We first form all possible pairs for the first list. In this case we would have the following pairs:

(1,2), (1,3), (1,4), (1,5), (1,6), (1,7), (1,8), (1,9), (1,10)

(2,3), (2,4), (2,5), (2,6), (2,7), (2,8), (2,9), (2,10)

(3,4), (3,5), (3,6), (3,7), (3,8), (3,9), (3,10)

(4,5), (4,6), (4,7), (4,8), (4,9), (4,10)

(5,6), (5,7), (5,8), (5,9), (5,10)

(6,7), (6,8), (6,9), (6,10)

(7,8), (7,9), (7,10)

(8,9), (8,10)

(9,10)

We proceed by forming all the pairs in the second list. In this case we would have the following pairs:

(1,2), (1,3), (1,4), (1,5)

(2,3), (2,4), (2,5)

(3,4), (3,5)

(4,5)

Using the first distance we get the following table:

Tab 1.6 List of distances

pair	distance in 1st list	distance in 2nd list
(1,2)	0.105885148	0.22352868
(1,3)	0.117643	0.22352871
(1,4)	0.117643	0.22352871
(1,5)	0.105885148	0.10588514
(2,3)	0.117643	0.22352871
(2,4)	0.117643	0.22352871
(2,5)	0.105885148	0.22352868
(3,4)	0.117643	0.22352869
(3,5)	0.117643	0.22352871
(4,5)	0.117643	0.22352871

In the case we are using the first distance we would look for the node where two points join. In the case we are using the second distance we would look for the number of tree edges joining those two points.

To compute the correlation coefficient we need to pick one pair in the second list and compute its distance and then look for the same pair in the first list and compute its distance. We do the same thing for all remaining pairs in the second list.

Examples of how trees are constructed are given in Fig 1.5, Fig 1.6, Fig 1.7. Following are all correlation coefficients computed. For each one of them there are 3*2*2 possible designs except for the last one where the third input is a combination of uniform distribution and the distribution (1.10).

first input { Slink, Clink, Average

second input { distance 1, distance 2

third input { uniform distribution, distribution (1.10)

1) First correlation coefficient: The first coefficient of correlation is the one between a list of pairs built from a sample S and a list of pairs built from the first half of the sample S. The first

half of S is taken before S is sorted.

2) Second correlation coefficient: The second coefficient of correlation is the one between a list of pairs built from a sample S and a list of pairs built from the second half of the sample S. The second half of S is taken before the sample S is sorted.

3) Third correlation coefficient: The third coefficient of correlation is the one between a list of pairs built from the first half of the sample S ,say Sfh, and a list of pairs built from the first half of another sample S', say Sfh'. The two samples are given the id's after being sorted. The first point of sample Sfh is given as id the number 1 and so is given the first point of sample Sfh'. The second point of sample Sfh is given as id the number 2 and so is given the second point of sample Sfh' and so on.

4) Fourth correlation coefficient: The fourth coefficient of correlation is the one between a list of pairs built from the second half of the sample S ,say Sfh, and a list of pairs built from the second half of another sample S', say Sfh'. The two samples are given the id's after being sorted. The first point of sample Sfh is given as id the number 1 and so is given the first point of sample Sfh'. The second point of sample Sfh is given as id the number 2 and so is given the second point of sample Sfh' and so on.

5) Fifth correlation coefficient: The fifth coefficient of correlation is the one between a list of pairs built from the sample S and a list of pairs built from the sample X added to the sample S. The sample X contains 10 new randomly generated points.

6) Sixth correlation coefficient: The sixth coefficient of correlation definition is the same as the fifth correlation coefficient except the sample X contains 20 new randomly generated points.

7) Seventh correlation coefficient: The seventh coefficient of correlation definition is the same as the fifth correlation coefficient except the sample X contains 30 new randomly generated points.

8) Eighth correlation coefficient: The eighth coefficient of correlation definition is the same as the fifth correlation coefficient except the sample X contains 40 new randomly generated points.

9) Ninth correlation coefficient: The ninth coefficient of correlation is the one between a list of pairs built from the sample S using the uniform distribution given by the formula (1.8) and a list of pairs built from the sample S' using the second distribution given by the formula (1.10). The sample X contains $10*i$ new randomly generated points where i is in [AUH83, Rom84].

We introduce some notations which are going to be used in Chapter 4. This is intended to give a readable version of the program output. We will give a shorthand names to indicate all possible program inputs. Following is a table showing the inputs with their shorthand indicator.

Tab 1.7 Terms shorthand

term	shorthand
Slink	S
Clink	C
Average	A
Unif distrib	U
Distrib (1.10)	O
Distance 1	1
Distance 2	2

In order to use a combination of terms we only need to concatenate the shorthands in the order showed above.

example: Suppose we want to have as input the following parameters:

1-Slink

2-Uniform distribution

3-Distance 1

We would indicate this input by the shorthand SU1.

Function Approximation of the Correlation Coefficients

We want to find the relationship between the correlation coefficients and the data size. We start by computing the average of 100 correlation coefficients taken between two lists computed from the same data size. We then compute the standard deviation of this average correlation coefficient using the formula (1.7). We proceed by performing the same steps for the other data sizes.

When we finish the two computations mentioned above we then proceed with the approximation of those points by a linear function

$$f(x) = ax + b \tag{1.12}$$

Using the least squares problem.

We now give the criterion of a good approximation: We say we have a good approximation if the inequation

$$| y_i - f(x_i) | \leq \sigma(y_i) \qquad \text{for all } i \tag{1.13}$$

, where y_i is the point to approximate, f is the approximation function, and $\sigma(y_i)$ is the standard deviation for y_i , is satisfied.

If the inequation (1.13) is not satisfied then we do not consider the linear function as a good approximation.

2. Results and their Interpretation

We start by giving some shorthands used in the subsequent tables:

The number n is the size of the data input.

The shordhand sd stands for the standard deviation.

The shorthand cc stands for the correlation coefficient.

The shorthand CDD stands for:

C: Clustering Method

D: Distribution function

D: Distance used

We now give some information about the entries of the two tables shown for each correlation coefficient:

Each entry (row) of the first table consists of two sub-entries:

a: The average of 100 correlation coefficients taken between two lists formed from the same kind of data sizes

b: The standard deviation of those 100 correlation coefficients.

It is easy to check that the inequation (1.13) is satisfied for all the correlation points. This is true for every function approximation shown in this chapter. Thus all approximations are good since they verify our criterion of goodness.

Tab 4.1 First correlation coefficient

n		SU1	SU2	CU1	CU2	AU1	AU2	SO1	SO2	CO1	CO2	AO1	AO2
10	cc	0.79	0.70	0.70	0.73	0.77	0.75	0.96	0.88	0.94	0.85	0.97	0.89
	sd	0.28	0.20	0.27	0.22	0.18	0.19	0.077	0.078	0.19	0.11	0.056	0.070
15	cc	0.75	0.65	0.72	0.72	0.76	0.74	0.95	0.87	0.89	0.83	0.97	0.89
	sd	0.34	0.32	0.23	0.17	0.17	0.16	0.090	0.076	0.15	0.18	0.065	0.067
20	cc	0.82	0.68	0.73	0.73	0.78	0.75	0.95	0.87	0.90	0.84	0.96	0.89
	sd	0.31	0.20	0.21	0.17	0.15	0.11	0.073	0.075	0.16	0.099	0.060	0.060
25	cc	0.77	0.64	0.73	0.73	0.73	0.73	0.95	0.87	0.88	0.84	0.95	0.89
	sd	0.28	0.23	0.20	0.15	0.24	0.16	0.075	0.071	0.17	0.13	0.059	0.064
30	cc	0.75	0.62	0.72	0.72	0.75	0.73	0.94	0.87	0.85	0.81s0	0.95	0.87
	sd	0.28	0.26	0.20	0.15	0.21	0.16	0.049	0.079	0.13	0.10	0.052	0.057
35	cc	0.68	0.59	0.77	0.74	0.76	0.75	0.94	0.86	0.85	0.82	0.95	0.87
	sd	0.28	0.19	0.16	0.15	0.16	0.14	0.059	0.068	0.17	0.10	0.13	0.053
40	cc	0.73	0.56	0.67	0.70	0.75	0.73	0.96	0.88	0.88	0.83	0.96	0.88
	sd	0.32	0.24	0.18	0.14	0.18	0.12	0.079	0.078	0.15	0.11	0.049	0.064
45	cc	0.74	0.57	0.70	0.69	0.71	0.71	0.96	0.86	0.86	0.81	0.96	0.87
	sd	0.25	0.22	0.21	0.12	0.19	0.12	0.11	0.067	0.17	0.11	0.054	0.053
50	cc	0.76	0.56	0.67	0.68	0.73	0.72	0.92	0.86	0.84	0.81	0.95	0.87
	sd	0.27	0.25	0.16	0.12	0.18	0.10	0.083	0.071	0.11	0.11	0.064	0.049

Tab 4.2 Function approximation

CDD	function approx
SU1	-0.00083 X + 0.81
SU2	-0.0018 X + 0.73
CU1	-0.00034 X + 0.73
CU2	-0.00020 X + 0.72
AU1	-0.00040 X + 0.77
AU2	-0.00011 X + 0.74
SO1	-0.00025 X + 0.96
SO2	-0.00007 X + 0.87
CO1	-0.0011 X + 0.95
CO2	-0.00051 X + 0.86
AO1	-0.00013 X + 0.97
AO2	-0.000073 X + 0.88

Tab 4.3 Second correlation coefficient

n		SU1	SU2	CU1	CU2	AU1	AU2	SO1	SO2	CO1	CO2	AO1	AO2
10	cc	0.78	0.74	0.79	0.77	0.81	0.78	0.97	0.88	0.91	0.84	0.94	0.87
	sd	0.22	0.22	0.16	0.12	0.19	0.12	0.094	0.092	0.15	0.12	0.13	0.098
15	cc	0.76	0.67	0.72	0.71	0.77	0.74	0.97	0.88	0.88	0.83	0.94	0.88
	sd	0.27	0.23	0.32	0.32	0.17	0.31	0.11	0.093	0.15	0.13	0.088	0.074
20	cc	0.84	0.70	0.73	0.75	0.79	0.78	0.95	0.87	0.90	0.85	0.95	0.88
	sd	0.26	0.23	0.23	0.18	0.18	0.10	0.089	0.088	0.16	0.11	0.078	0.072
25	cc	0.80	0.68	0.72	0.74	0.76	0.76	0.94	0.87	0.88	0.85	0.94	0.88
	sd	0.31	0.23	0.23	0.18	0.18	0.11	0.050	0.077	0.16	0.11	0.097	0.072
30	cc	0.79	0.65	0.72	0.72	0.78	0.76	0.95	0.86	0.88	0.83	0.94	0.87
	sd	0.28	0.21	0.21	0.15	0.17	0.11	0.049	0.067	0.16	0.094	0.080	0.052
35	cc	0.69	0.62	0.71	0.70	0.76	0.75	0.94	0.85	0.85	0.81	0.94	0.87
	sd	0.25	0.22	0.20	0.15	0.16	0.17	0.082	0.081	0.18	0.090	0.079	0.049
40	cc	0.72	0.57	0.67	0.69	0.76	0.74	0.95	0.88	0.87	0.83	0.97	0.89
	sd	0.27	0.26	0.19	0.15	0.15	0.15	0.092	0.085	0.16	0.10	0.068	0.054
45	cc	0.74	0.55	0.69	0.69	0.74	0.71	0.95	0.87	0.85	0.81	0.95	0.87
	sd	0.31	0.20	0.18	0.14	0.27	0.12	0.11	0.070	0.15	0.10	0.056	0.058
50	cc	0.77	0.55	0.71	0.70	0.73	0.73	0.92	0.86	0.82	0.81	0.94	0.87
	sd	0.27	0.29	0.18	0.13	0.22	0.13	0.082	0.076	0.17	0.096	0.064	0.058

Tab 4.4 Function approximation

CDD	function approx
SU1	-0.0012 X + 0.84
SU2	-0.0024 X + 0.78
CU1	-0.0012 X + 0.80
CU2	-0.00082 X + 0.77
AU1	-0.0011 X + 0.84
AU2	-0.00052 X + 0.78
SO1	-0.00016 X + 0.95
SO2	0.00017 X + 0.85
CO1	-0.00089 X + 0.93
CO2	-0.00030 X + 0.85
AO1	0.000063 X + 0.94
AO2	0.00022 X + 0.86

Tab 4.5 Third correlation coefficient

n		SU1	SU2	CU1	CU2	AU1	AU2	SO1	SO2	CO1	CO2	AO1	AO2
10	cc	0.32	0.41	0.50	0.49	0.53	0.51	0.46	0.57	0.56	0.60	0.53	0.61
	sd	0.18	0.20	0.22	0.20	0.22	0.20	0.34	0.22	0.30	0.19	0.31	0.19
15	cc	0.35	0.42	0.55	0.54	0.58	0.57	0.47	0.56	0.59	0.61	0.56	0.61
	sd	0.18	0.17	0.19	0.17	0.20	0.17	0.30	0.23	0.24	0.16	0.27	0.17
20	cc	0.32	0.40	0.60	0.57	0.60	0.57	0.52	0.58	0.67	0.65	0.63	0.66
	sd	0.17	0.18	0.18	0.13	0.14	0.10	0.29	0.27	0.22	0.13	0.26	0.17
25	cc	0.33	0.42	0.61	0.58	0.64	0.61	0.53	0.57	0.66	0.64	0.63	0.65
	sd	0.17	0.14	0.14	0.11	0.14	0.095	0.26	0.24	0.18	0.17	0.24	0.17
30	cc	0.38	0.41	0.61	0.57	0.64	0.60	0.60	0.63	0.66	0.64	0.68	0.67
	sd	0.14	0.16	0.19	0.14	0.14	0.11	0.22	0.17	0.18	0.15	0.19	0.15
35	cc	0.36	0.41	0.60	0.59	0.63	0.61	0.61	0.63	0.68	0.65	0.69	0.67
	sd	0.16	0.14	0.16	0.096	0.13	0.087	0.21	0.16	0.16	0.11	0.18	0.15
40	cc	0.38	0.41	0.63	0.60	0.67	0.62	0.60	0.63	0.64	0.62	0.66	0.66
	sd	0.14	0.15	0.17	0.092	0.14	0.086	0.23	0.16	0.26	0.12	0.16	0.11
45	cc	0.34	0.42	0.63	0.58	0.66	0.62	0.63	0.65	0.71	0.66	0.71	0.68
	sd	0.13	0.14	0.15	0.11	0.14	0.097	0.23	0.13	0.16	0.10	0.16	0.11
50	cc	0.36	0.43	0.61	0.59	0.66	0.63	0.65	0.65	0.73	0.68	0.74	0.69
	sd	0.14	0.15	0.15	0.091	0.14	0.089	0.23	0.16	0.17	0.10	0.16	0.11

Tab 4.6 Function approximation

CDD	function approx
SU1	0.00072 X + 0.30
SU2	0.00048 X + 0.38
CU1	0.0017 X + 0.49
CU2	0.0016 X + 0.47
AU1	0.0023 X + 0.47
AU2	0.0020 X + 0.46
SO1	0.0030 X + 0.38
SO2	0.0018 X + 0.49
CO1	0.0023 X + 0.51
CO2	0.0014 X + 0.55
AO1	0.0030 X + 0.46
AO2	0.0017 X + 0.54

Tab 4.7 Fourth correlation coefficient

n		SU1	SU2	CU1	CU2	AU1	AU2	SO1	SO2	CO1	CO2	AO1	AO2
10	cc	0.25	0.37	0.51	0.50	0.48	0.49	0.47	0.54	0.58	0.57	0.55	0.57
	sd	0.46	0.19	0.23	0.19	0.22	0.19	0.26	0.25	0.27	0.24	0.26	0.25
15	cc	0.31	0.41	0.58	0.55	0.56	0.55	0.45	0.54	0.63	0.61	0.58	0.60
	sd	0.14	0.18	0.18	0.15	0.17	0.14	0.27	0.22	0.22	0.17	0.19	0.21
20	cc	0.35	0.44	0.57	0.55	0.59	0.58	0.54	0.58	0.66	0.63	0.65	0.63
	sd	0.15	0.15	0.17	0.12	0.18	0.13	0.25	0.22	0.20	0.16	0.19	0.18
25	cc	0.37	0.45	0.63	0.60	0.65	0.60	0.56	0.58	0.70	0.65	0.66	0.65
	sd	0.16	0.15	0.17	0.12	0.13	0.10	0.24	0.21	0.17	0.12	0.14	0.16
30	cc	0.36	0.44	0.62	0.59	0.65	0.60	0.61	0.65	0.72	0.66	0.70	0.69
	sd	0.15	0.15	0.17	0.11	0.13	0.11	0.24	0.16	0.18	0.10	0.10	0.16
35	cc	0.37	0.42	0.62	0.60	0.64	0.61	0.61	0.63	0.74	0.68	0.71	0.69
	sd	0.14	0.15	0.17	0.10	0.17	0.084	0.22	0.18	0.17	0.11	0.11	0.16
40	cc	0.38	0.42	0.62	0.59	0.65	0.62	0.66	0.67	0.69	0.67	0.73	0.70
	sd	0.15	0.14	0.15	0.097	0.15	0.088	0.24	0.15	0.16	0.11	0.11	0.18
45	cc	0.34	0.41	0.65	0.59	0.68	0.62	0.64	0.64	0.67	0.65	0.71	0.68
	sd	0.21	0.15	0.14	0.10	0.15	0.10	0.20	0.12	0.16	0.13	0.18	0.12
50	cc	0.34	0.39	0.64	0.60	0.66	0.62	0.63	0.65	0.71	0.67	0.73	0.69
	sd	0.16	0.15	0.16	0.096	0.14	0.086	0.23	0.15	0.16	0.093	0.15	0.095

Tab 4.8 Function approximation

CDD	function approx
SU1	0.00097 X + 0.28
SU2	0.00056 X + 0.38
CU1	0.0018 X + 0.49
CU2	0.0016 X + 0.47
AU1	0.0024 X + 0.47
AU2	0.0020 X + 0.46
SO1	0.0034 X + 0.36
SO2	0.0027 X + 0.43
CO1	0.0027 X + 0.49
CO2	0.0025 X + 0.48
AO1	0.0035 X + 0.44
AO2	0.0029 X + 0.47

Tab 4.9 Fifth correlation coefficient

n		SU1	SU2	CU1	CU2	AU1	AU2	SO1	SO2	CO1	CO2	AO1	AO2
10	cc	0.68	0.62	0.75	0.69	0.79	0.69	0.97	0.83	0.93	0.83	0.96	0.83
	sd	0.30	0.25	0.19	0.23	0.17	0.21	0.059	0.12	0.13	0.13	0.081	0.14
20	cc	0.83	0.69	0.71	0.74	0.76	0.73	0.98	0.96	0.92	0.85	0.96	0.87
	sd	0.20	0.17	0.27	0.14	0.16	0.12	0.050	0.093	0.14	0.11	0.060	0.087
30	cc	0.85	0.72	0.78	0.80	0.78	0.77	0.99	0.89	0.90	0.86	0.97	0.89
	sd	0.19	0.16	0.17	0.12	0.16	0.11	0.030	0.058	0.16	0.11	0.045	0.076
40	cc	0.89	0.75	0.73	0.77	0.76	0.76	0.98	0.90	0.89	0.87	0.96	0.89
	sd	0.16	0.12	0.21	0.11	0.16	0.11	0.047	0.053	0.15	0.099	0.051	0.062
50	cc	0.89	0.77	0.74	0.78	0.76	0.76	0.98	0.90	0.89	0.87	0.96	0.90
	sd	0.15	0.11	0.19	0.11	0.16	0.10	0.033	0.051	0.15	0.082	0.050	0.066
60	cc	0.88	0.76	0.74	0.77	0.74	0.76	0.99	0.92	0.90	0.88	0.97	0.90
	sd	0.19	0.12	0.18	0.10	0.24	0.098	0.031	0.038	0.13	0.069	0.042	0.043
70	cc	0.92	0.77	0.79	0.80	0.75	0.76	0.99	0.92	0.90	0.89	0.97	0.90
	sd	0.081	0.11	0.17	0.10	0.18	0.094	0.024	0.040	0.15	0.079	0.041	0.043
80	cc	0.90	0.77	0.79	0.81	0.75	0.77	0.99	0.93	0.89	0.89	0.97	0.91
	sd	0.17	0.11	0.18	0.10	0.14	0.098	0.017	0.047	0.14	0.078	0.036	0.048
90	cc	0.95	0.80	0.75	0.80	0.76	0.78	0.99	0.93	0.89	0.89	0.97	0.90
	sd	0.056	0.095	0.17	0.10	0.14	0.081	0.021	0.047	0.15	0.083	0.038	0.053
100	cc	0.96	0.82	0.78	0.82	0.73	0.77	0.99	0.94	0.86	0.88	0.95	0.89
	sd	0.096	0.11	0.19	0.11	0.16	0.091	0.018	0.037	0.16	0.090	0.048	0.044

Tab 4.10 Function approximation

CDD	function approx
SU1	0.0022 X + 0.75
SU2	0.0018 X + 0.65
CU1	0.00044 X + 0.73
CU2	0.0011 X + 0.72
AU1	-0.00042 X + 0.78
AU2	0.00061 X + 0.72
SO1	0.00022 X + 0.97
SO2	0.0011 X + 0.84
CO1	-0.00051 X + 0.92
CO2	0.00053 X + 0.84
AO1	-0.0000092 X + 0.97
AO2	0.00056 X + 0.86

Tab 4.11 Sixth correlation coefficient

n		SU1	SU2	CU1	CU2	AU1	AU2	SO1	SO2	CO1	CO2	AO1	AO2
10	cc	0.68	0.60	0.72	0.65	0.78	0.67	0.95	0.80	0.92	0.78	0.97	0.80
	sd	0.30	0.23	0.27	0.23	0.17	0.23	0.092	0.14	0.16	0.14	0.070	0.14
20	cc	0.77	0.67	0.73	0.71	0.78	0.72	0.96	0.82	0.90	0.81	0.95	0.94
	sd	0.22	0.18	0.21	0.14	0.16	0.14	0.075	0.11	0.13	0.13	0.056	0.11
30	cc	0.80	0.66	0.69	0.72	0.77	0.74	0.97	0.83	0.89	0.82	0.96	0.85
	sd	0.20	0.19	0.19	0.12	0.16	0.11	0.051	0.082	0.14	0.11	0.056	0.097
40	cc	0.82	0.66	0.73	0.74	0.77	0.75	0.97	0.85	0.89	0.84	0.96	0.87
	sd	0.19	0.14	0.17	0.10	0.16	0.11	0.045	0.064	0.15	0.091	0.046	0.068
50	cc	0.84	0.67	0.71	0.73	0.76	0.74	0.97	0.86	0.88	0.85	0.96	0.88
	sd	0.16	0.14	0.17	0.11	0.15	0.11	0.036	0.066	0.16	0.089	0.045	0.055
60	cc	0.86	0.68	0.70	0.71	0.76	0.73	0.97	0.88	0.86	0.84	0.96	0.88
	sd	0.16	0.12	0.16	0.10	0.14	0.10	0.032	0.048	0.15	0.090	0.045	0.047
70	cc	0.89	0.69	0.70	0.72	0.73	0.72	0.98	0.87	0.84	0.84	0.96	0.87
	sd	0.083	0.12	0.17	0.11	0.25	0.099	0.029	0.049	0.14	0.075	0.041	0.042
80	cc	0.87	0.69	0.71	0.75	0.75	0.74	0.98	0.90	0.88	0.86	0.97	0.89
	sd	0.15	0.15	0.17	0.10	0.14	0.091	0.024	0.045	0.15	0.081	0.044	0.053
90	cc	0.89	0.70	0.72	0.74	0.75	0.74	0.99	0.90	0.85	0.85	0.96	0.89
	sd	0.13	0.12	0.15	0.085	0.19	0.074	0.029	0.055	0.16	0.081	0.050	0.049
100	cc	0.90	0.70	0.73	0.76	0.71	0.73	0.99	0.90	0.84	0.84	0.95	0.87
	sd	0.15	0.14	0.17	0.094	0.17	0.088	0.016	0.048	0.15	0.083	0.072	0.048

Tab 4.12 Function approximation

CDD	function approx
SU1	0.0021 X + 0.72
SU2	0.00084 X + 0.63
CU1	0.000065 X + 0.71
CU2	0.00080 X + 0.68
AU1	-0.00063 X + 0.79
AU2	0.00031 X + 0.71
SO1	0.00046 X + 0.95
SO2	0.0012 X + 0.80
CO1	-0.00076 X + 0.92
CO2	0.00055 X + 0.80
AO1	-0.000015 X + 0.96
AO2	0.00070 X + 0.82

Tab 4.13 Seventh correlation coefficient

n		SU1	SU2	CU1	CU2	AU1	AU2	SO1	SO2	CO1	CO2	AO1	AO2
10	cc	0.62	0.56	0.70	0.62	0.74	0.62	0.94	0.80	0.91	0.76	0.97	0.80
	sd	0.32	0.22	0.26	0.20	0.26	0.21	0.084	0.012	0.15	0.16	0.058	0.14
20	cc	0.75	0.62	0.70	0.69	0.77	0.70	0.94	0.81	0.89	0.79	0.96	0.83
	sd	0.23	0.18	0.21	0.14	0.15	0.14	0.083	0.10	0.17	0.12	0.057	0.11
30	cc	0.77	0.60	0.72	0.70	0.77	0.70	0.94	0.82	0.84	0.79	0.95	0.85
	sd	0.21	0.18	0.20	0.14	0.16	0.14	0.078	0.097	0.14	0.13	0.055	0.087
40	cc	0.80	0.62	0.71	0.71	0.76	0.72	0.96	0.83	0.87	0.82	0.95	0.85
	sd	0.19	0.16	0.16	0.11	0.15	0.11	0.066	0.090	0.15	0.093	0.048	0.077
50	cc	0.81	0.63	0.70	0.70	0.74	0.71	0.97	0.84	0.87	0.82	0.95	0.85
	sd	0.17	0.15	0.17	0.11	0.18	0.11	0.040	0.078	0.15	0.083	0.060	0.070
60	cc	0.80	0.63	0.72	0.70	0.74	0.71	0.97	0.84	0.84	0.80	0.96	0.85
	sd	0.17	0.16	0.17	0.11	0.18	0.10	0.043	0.069	0.16	0.085	0.047	0.061
70	cc	0.85	0.64	0.72	0.71	0.73	0.71	0.97	0.84	0.81	0.80	0.96	0.86
	sd	0.13	0.13	0.17	0.11	0.17	0.093	0.050	0.071	0.20	0.084	0.042	0.046
80	cc	0.83	0.63	0.70	0.71	0.76	0.72	0.98	0.86	0.87	0.84	0.97	0.87
	sd	0.16	0.12	0.15	0.094	0.13	0.091	0.038	0.076	0.16	0.090	0.053	0.059
90	cc	0.86	0.66	0.69	0.70	0.76	0.73	0.98	0.88	0.85	0.83	0.95	0.87
	sd	0.14	0.12	0.15	0.082	0.15	0.072	0.035	0.063	0.15	0.091	0.048	0.055
100	cc	0.87	0.67	0.72	0.73	0.74	0.73	0.98	0.87	0.82	0.81	0.95	0.86
	sd	0.16	0.12	0.16	0.094	0.16	0.098	0.038	0.062	0.15	0.087	0.066	0.048

Tab 4.14 Function approximation

CDD	function approx
SU1	0.0021 X + 0.68
SU2	0.00085 X + 0.58
CU1	0.00004 X + 0.71
CU2	0.00069 X + 0.66
AU1	-0.00014 X + 0.76
AU2	0.00076 X + 0.66
SO1	0.00049 X + 0.93
SO2	0.00082 X + 0.79
CO1	-0.00071 X + 0.90
CO2	0.00052 X + 0.78
AO1	-0.00011X + 0.96
AO2	0.00060 X + 0.82

Tab 4.15 Eighth correlation coefficient

n		SU1	SU2	CU1	CU2	AU1	AU2	SO1	SO2	CO1	CO2	AO1	AO2
10	cc	0.64	0.55	0.69	0.63	0.74	0.65	0.92	0.80	0.88	0.77	0.97	0.81
	sd	0.31	0.26	0.29	0.22	0.27	0.23	0.097	0.12	0.17	0.16	0.055	0.13
20	cc	0.77	0.58	0.71	0.67	0.75	0.69	0.92	0.81	0.86	0.78	0.95	0.83
	sd	0.23	0.18	0.20	0.14	0.15	0.13	0.11	0.12	0.15	0.11	0.056	0.11
30	cc	0.81	0.60	0.72	0.71	0.76	0.71	0.94	0.82	0.86	0.80	0.96	0.86
	sd	0.19	0.17	0.20	0.12	0.15	0.11	0.11	0.092	0.15	0.12	0.050	0.083
40	cc	0.82	0.62	0.69	0.68	0.74	0.71	0.94	0.82	0.84	0.80	0.94	0.85
	sd	0.18	0.14	0.17	0.12	0.20	0.10	0.099	0.082	0.18	0.12	0.068	0.063
50	cc	0.79	0.60	0.70	0.68	0.75	0.71	0.95	0.84	0.86	0.82	0.94	0.85
	sd	0.18	0.17	0.18	0.12	0.15	0.11	0.083	0.077	0.16	0.083	0.056	0.050
60	cc	0.75	0.60	0.71	0.69	0.74	0.71	0.97	0.82	0.84	0.80	0.95	0.84
	sd	0.20	0.17	0.16	0.12	0.16	0.099	0.040	0.061	0.15	0.088	0.046	0.088
70	cc	0.79	0.58	0.68	0.68	0.74	0.71	0.97	0.83	0.78	0.78	0.95	0.85
	sd	0.16	0.16	0.16	0.11	0.16	0.092	0.037	0.056	0.17	0.085	0.045	0.041
80	cc	0.81	0.60	0.72	0.70	0.76	0.72	0.98	0.85	0.85	0.81	0.96	0.87
	sd	0.18	0.14	0.15	0.099	0.15	0.095	0.041	0.067	0.18	0.085	0.046	0.054
90	cc	0.83	0.61	0.71	0.71	0.74	0.73	0.98	0.85	0.83	0.81	0.95	0.87
	sd	0.13	0.15	0.15	0.10	0.17	0.091	0.038	0.062	0.15	0.89	0.053	0.053
100	cc	0.86	0.61	0.70	0.71	0.74	0.73	0.96	0.85	0.81	0.80	0.94	0.87
	sd	0.13	0.14	0.15	0.11	0.16	0.094	0.070	0.059	0.15	0.092	0.056	0.048

Tab 4.16 Function approximation

CDD	function approx
SU1	0.0014 X + 0.71
SU2	0.00032 X + 0.58
CU1	0.00009 X + 0.70
CU2	0.00059 X + 0.65
AU1	-0.000073 X + 0.75
AU2	0.00064 X + 0.67
SO1	0.00063 X + 0.92
SO2	0.00054 X + 0.80
CO1	-0.00063 X + 0.88
CO2	0.00029 X + 0.78
AO1	-0.00012 X + 0.96
AO2	0.00053 X + 0.82

Tab 4.17 Ninth correlation coefficient

n		SUO1	SUO2	CUO1	CUO2	AUO1	AUO2
10	cc	0.33	0.45	0.51	0.52	0.49	0.54
	sd	0.15	0.27	0.25	0.22	0.25	0.22
15	cc	0.31	0.46	0.54	0.54	0.53	0.54
	sd	0.13	0.14	0.21	0.15	0.21	0.16
20	cc	0.29	0.42	0.56	0.57	0.53	0.55
	sd	0.12	0.20	0.18	0.12	0.18	0.13
25	cc	0.26	0.43	0.58	0.57	0.53	0.54
	sd	0.13	0.20	0.17	0.10	0.17	0.14
30	cc	0.27	0.42	0.59	0.57	0.54	0.54
	sd	0.12	0.17	0.17	0.12	0.17	0.11
35	cc	0.25	0.42	0.60	0.57	0.57	0.56
	sd	0.28	0.17	0.15	0.12	0.16	0.12
40	cc	0.24	0.40	0.63	0.58	0.57	0.56
	sd	0.19	0.16	0.18	0.12	0.17	0.12
45	cc	0.24	0.40	0.59	0.58	0.57	0.59
	sd	0.19	0.18	0.16	0.10	0.15	0.093
50	cc	0.22	0.40	0.60	0.59	0.60	0.62
	sd	0.14	0.15	0.15	0.086	0.15	0.090

Tab 4.18 Function approximation

CDD	function approx
SUO1	-0.0023 X + 0.34
SUO2	-0.00093 X + 0.45
CUO1	0.0020 X + 0.52
CUO2	0.0013 X + 0.52
AUO1	0.0021 X + 0.48
AUO2	0.0016 X + 0.51

a) First category of correlation coefficients: The first, second, third, and fourth correlation coefficients are meant to check the influence of the context upon the way points are clustered.

b) Second category of correlation coefficients: The fifth, sixth, seventh, and eight correlation coefficients are meant to check the influence of the context size upon the way points are clustered.

c) Third category of correlation coefficients: The ninth correlation coefficient is meant to check the relation which might exist between two lists of pairs formed from two samples which are generated using different distributions.

We conducted the same kind of experiment using the step size equal to five (5) and ten (10) to see whether the results are going to be sensible to the step size or not. Those results using a step size equal to five (5) do not differ in any way from the results obtained using the step size equal to ten (10). This fact merely confirms the conclusions drawn from the tables (1.1) to (1.18).

3. Conclusion

In the past, some comparative studies have been conducted and narrowed to a specific set of input [Boy69] while others have been dealing with the theoretical aspects of the methods used in Cluster Analysis [Sne69]. Our work is the beginning of a more general comparison among some clustering methods since numerous criteria are used to carry out the comparison.

1) First observation: The results shown in tables 1.1, 1.2, 1.3, and 1.4 show that the context does not completely *hide* the samples. Indeed, we notice that the correlation coefficients shown in tables 1.1 and 1.2 are quite different than the ones shown in tables 1.3 and 1.4. This fact clearly shows what correlation coefficient we are computing.

2) Second observation: The tables 1.1 to 1.4 and the tables 1.9 to 1.18 clearly demonstrate that the context does not influence the way the data are clustered since all the correlation coefficients are close to one (1). Moreover, those results demonstrate that the three (3) methods are equally stable. This fact is rather surprising since we expect that the average method would be more stable than the other clustering methods.

3) third observation: The third and fourth correlation coefficient compared to the ninth correlation coefficient surprisingly show that the distribution does not matter when clustering a set of points. Indeed, the correlation coefficients shown in the respective tables are sensibly the same. One expects that two lists taken from the same distribution would be more similar than the ones taken from different distributions.

4) Fourth observation: The taxonomic distance does not influence the way the correlation coefficients are computed since there is no significant difference between the correlation coefficients using the first distance and those using the second distance. This fact is a constant

for all tables shown in chapter 4.

5) Fifth observation: The tables 1.17 and 1.18 also show that the three (3) methods would be good even in a noisy environment since there is no difference of behavior with the results obtained when using the uniform distribution.

6) Sixth observation: The last observation has to do with the fact that all tables, clearly show that no method shows to be better than the others no matter what criterion we used. However, Slink and Clink methods are computationally more attractive than Average method.

Some conclusions drawn above are rather surprising since the results conflict with the intuition we had before undertaking this study. For future work we would recommend developing a program having as a working set the multi-dimensional space. We also recommend working with other kinds of distribution of random numbers. We expect that the more dimensions we use as working set the closer the correlation coefficient of two lists to zero (0) would be. This is our expectation and it would be interesting to see whether the experiments will confirm or infirm our expectation. This would give a more general comparison and conclusion among the three (3) methods mentioned above.

Acknowledgment

Andrzej Euhrenfeucht has been instrumental in the research presented in this paper. We would like to acknowledge his valuable contribution.

Reference

[AUH83] A. V. Aho, J. D. Ullman and J. E. Hopcroft, *Data Structure and Algorithms*, Addison-Wesley, Reading, Mass, 1983.

[Bou87] A. Bouguettaya, *A Comparative Study of Some Clustering Methods with On-line Data*, University of Colorado At Boulder, Boulder, Colorado, July, 1987. MS Thesis in Computer Science.

[Boy69] H. C. Boyce, ''Mapping Diversity: A Comparative Study of Some Numerical Methods'', *Numerical Taxonomy, Proceedings of the Colloquium in Numerical Taxonomy*, London, 1969.

[HuK88] S. E. Hudson and R. King, ''CACTIS: A Self-Adaptive, Concurrent Implementation of an Object-Oriented Database Management System'', *Trans. Database Systems 14[3]* (1988), 291-321.

[Knu71] D. E. Knuth, *The Art of Computer Programming Vol.2: Seminumerical Algorithms*, Addison-Wesley, Reading, Mass, 1971.

[Rom84] H. C. Romesburg, *Cluster Analysis for Researchers*, Lifetime Learning Publications, London, 1984.

[SnS63] P. Sneath and R. S. Sokal, *Principles of Numeriacl Taxonomy*, W. H. Freeman, San Francisco, 1963.

[Sne69] P. Sneath, ''Evaluation of Clustering Methods '', *Numerical Taxonomy, Proceedings of the Colloquium in Numerical Taxonomy*, London, 1969, 257-267.

[SnS73] P. Sneath and R. S. Sokal, ''Numerical Taxonomy'', *Principles and practice of Numerical Classification*, San Francisco, 1973.

[TsN91] M. M. Tsangaris and J. F. Naughton, ''A Stochastic Approach for Clustering in Object Bases'', *SIGMOD*, Denver, Colorado, May 1991.