DATA REDUCTION

TWO-DIMENSIONAL DATA CLUSTERING USING GROUP-BASED DISTANCES

Data clustering enjoys wide application in diverse fields such as data mining, access structures, knowledge discovery, software engineering, organization of information systems, and machine learning. In this article, the behavior and stability of two clustering techniques are examined: unweighted pair-group using arithmetic averages and Ward clustering. Three different statistical distributions are used to express how data objects are drawn from a two-dimensional space. In addition, two types of distances are utilized to compare the resulting trees: Euclidean and Edge distances. The results of an exhaustive set of experiments that involve data derived from two-dimensional spaces are presented. These experiments indicate a surprisingly high level of similarity between the two methods under most combinations of parameter settings.

The main objective of cluster analysis is to create groups of objects based on the degree of their association (1,2). Similarities among otherwise distinct data objects are exploited so that these objects are classified into groups. Cluster analysis has been used to determine taxonomy relationships among entities in diverse disciplines including management and classification of species (1), derivation of medical profiles (2,3), census and survey problems (4), databases (5), information retrieval (6), software engineering (7,8) as well as machine learning and data compression (9).

In database clustering, the ability to categorize data objects into groups allows the reallocation of related data to improve the performance of DBMSs. Good placement of objects could significantly decrease the response time needed to query object-oriented databases (OODBs) (5) and help further improve the performance of relational systems (10). Data records which are frequently referenced together are moved in close proximity to reduce access time. To reach this goal, cluster analysis is used to form clusters based on the similarities of data objects. Data may be reallocated based on values of an attribute, group of attributes, or on accessing patterns. By reallocating data objects, related records are physically placed closely together. These criteria determine the measuring *distance* among data objects. Hence, it is anticipated that the number of disk accesses required to obtain required data for the materialization of queries will diminish.

With the proliferation of OODBs the need for good performance clustering techniques becomes more crucial if acceptable overall performance is to be maintained. Some OODBs have already incorporated clustering strategies to improve query response times; however, these strategies are mostly heuristic and static in nature (11). The case of OODBs is unique in that the underlying model provides a testbed for dynamic clustering. Recently, a number of studies have appeared dealing with this problem (12,13,5,14,15). In addition, there have been studies that investigate adaptive clustering techniques. In this context, clustering techniques can effectively cope with changing access pattern and perform on-line grouping (16,10). The need for data clustering becomes even more pressing in light of contemporary systems and applications such as distributed databases, data mining, and knowledge discovery. Frequently in distributed databases voluminous data unable to be stored in a single site are fragmented and dispersed in a number of remote sites (17). If requested and unclustered data are located at different locations they can have tremendous impact on distributed query response times. In data mining and knowledge discovery (18,19), cluster analysis can be used to reveal data associations not previously encountered (20).

We use the term "objects" in a broad sense. They can be anything that requires classification based on a number of criteria. For instance, objects can represent attributes in relational databases (13), complex objects in object-oriented systems (15), software modules (7,8), etc. The only hard requirement needed is that they can be mapped as a unique point in a measurement space. Obviously, all objects to be clustered should be defined in the same measurement space. The way to evaluate the degree of similarities among a number of objects to be clustered varies according to the application domain and the characteristics of data used. Most of the work done today addresses problems where objects are mapped as points in one dimensional environment (21,15,7,5,14,8). More specifically, objects are represented as points belonging to a segment defined by an interval $[a, b]$ where a and b are arbitrary numbers.

In this article, we carry out an exhaustive study of known clustering techniques involving objects in the two-dimensional space. This type of data objects is pervasive to spatial databases, image databases, and so on (22). Multidimensional indexing techniques and temporal databases (23) may also tremendously benefit from efficient clustering analysis techniques. There has been little reported work evaluating clustering in the above context. In this study, our aim is to investigate the impact of two-dimension objects generation on the clustering process. Issues examined include:

- Calculation of the degree of association between different types of data.
- Determination of an acceptable criterion to evaluate the "quality" of clustering methods.
- Adaptability of the clustering methods with different distributions of data: uniformly distributed, skewed or concentrated around certain regions, etc.

The work reported here builds upon previous work that we have conducted using clustering algorithms such as *Slink, Clink,* and *Average* in the one-dimensional space (16). Our experimental framework takes into consideration a variety of environment parameters in order to test the clustering techniques sensibility and behavior.

The organization of the article is as follows. In the first section, the clustering methods used in this study are described. Following that, we detail the experiments conducted in this study, provide the interpretations of the experiment results, and finally offer some concluding remarks.

CLUSTER ANALYSIS METHODS

Groups of Objects and Distances

Cluster analysis groups entities that comply with a set of definitions (rules). A formed group should include objects that demonstrate very high degree of association. Hence, a cluster can be viewed as a group of *similar* or resembling objects. The primary goal of clustering is to produce homogeneous entities. Homogeneity refers to the common properties of the objects to be clustered. In addition, clustering displays, summarizes, predicts, and provides a basis for understanding patterns of behavior. Clusters of objects are displayed so that differences and similarities become apparent at a glance. Properties of clusters are highlighted by hiding properties of individuals. Thus, clusters easily isolated offer a basis for understanding, and speculations can be derived about the structure of the cluster system. Unusual (or unexpected) formulations may reveal anomalies that need special consideration and attention.

Clusters can be represented in the measurement space in the same way as the objects they contain. From that point of view, a single point is a cluster containing exactly one object. There are generally two ways to represent clusters in a measurement space as:

- a hypothetical point which is not an object in the cluster, or as
- an existing object in the cluster called centroid or cluster representative.

To cluster data objects in a database system or in any other environment, some means of quantifying the degree of associations between items is needed. This can be a measure of distances or similarities. There is a number of similarity measures available and the choice may have an effect on the results obtained. Multidimensional objects may use relative or normalized weight to convert their distance to an arbitrary scale so they can be compared. Once the objects are defined in the same measurement space as the points, it is then possible to compute the degree of similarity. In this respect, the smaller the distance the more similar two objects are. The most popular choice in computing distance is the *Euclidean distance* with:

$$d(i,j) = \sqrt{(x_{i_1} - x_{j_1})^2 + (x_{i_2} - x_{j_2})^2 + \cdots + (x_{i_n} - x_{j_n})^2} \quad (1)$$

where n is the number of dimensions. Consequently for the one-dimensional space, the distance becomes:

$$d(i,j) = |x_i - x_j| \quad (2)$$

Coefficients of correlation are the measurement that describe the strength of the relationship between two variables $\mathscr{X}$ and $\mathscr{Y}$. It essentially answers the question *how similar are* $\mathscr{X}$ and $\mathscr{Y}$?. The values of the coefficients of correlation range from 0 to 1 where the value 0 points to *no similarity* and the value 1 points to *high similarity.* The coefficient of correlation is used to find the similarity among (clustering) objects. The correlation r of two random variables $\mathscr{X}$ and $\mathscr{Y}$ where: $\mathscr{X} = (x_1, x_2, x_3, \ldots, x_n)$ and $\mathscr{Y} = (y_1, y_2, y_3, \ldots, y_n)$ is given by the formula:

$$r = \frac{|E(\mathscr{X},\mathscr{Y}) - E(\mathscr{X}) \cdot E(\mathscr{Y})|}{\sqrt{(E(\mathscr{X}^2) - E^2(\mathscr{X})}\sqrt{(E(\mathscr{Y}^2) - E^2(\mathscr{Y})}} \quad (3)$$

where $E(\mathscr{X}) = (\sum_{i=1}^{n} x_i)/n$, $E(\mathscr{Y}) = (\sum_{i=1}^{n} y_i)/n$, and $E(\mathscr{X}, \mathscr{Y}) = (\sum_{i=1}^{n} x_i \cdot y_i)/n$

Methods of Clustering

Clustering methods can be classified according to the type of the group structures they produce: partitioning or hierarchical.

The first family is widely used and methods here divide a given data set of N objects into M clusters with no overlapping allowed. These algorithms are known as partitioning methods. Here, a cluster may be represented by a *centroid* or *cluster representative* that represents the characteristics of all contained objects. It should be noted that this method is predominantly based on heuristics.

On the other hand, hierarchical methods work mostly in a bottom-up or top-down fashion. In the example of the bottom-up approach, the algorithm proceeds by performing a series of successive fusions. This produces a nested data set in which pairs of items or clusters are successively linked until every item in the data set is linked to form one cluster. Hierarchical methods can be further categorized as:

- Agglomerative in which $N-1$ pairwise joins are produced from an unclustered data set. In other words, from N clusters of one object, this method gradually forms one cluster of N objects. At each step, clusters or objects are

joined together into larger clusters ending with one big cluster containing all objects.
- Divisive in which all objects belong to a single cluster at the beginning, then they are divided into smaller clusters until the last cluster containing two objects have been broken apart into atomic constituents.

In both families of methods, the result of the procedure is a hierarchical tree. This tree is often presented as a *dendrogram,* in which pairwise couplings of the objects in the data set are shown and the length of the branches (vertices) or the value of the similarity is expressed numerically. Divisive methods are less commonly used (24) and in this article, we only discuss agglomerative techniques. As we are targeting the area of databases, agglomerative approaches naturally fit in within this paradigm (13,14,11).

Clustering Techniques

In this section, we discuss hierarchical agglomerative clustering methods and their characteristics. More specifically, we focus on two methods that enjoy wide usage (1,25).

Group Average Link Method. This method uses the average values pairwise distance, denoted $\mathcal{D}_{X,Y}$, within each participating cluster to determine similarity. All participating objects contribute to intercluster similarity. There are two different submethods based on this approach: Unweighted Pair-Group using Arithmetic Averages (UPGMA) and Weighted Pair-Group using Arithmetic Averages (WPGMA). The WPGMA is a special case of UPGMA. In WPGMA, the smaller cluster is leveled with the larger one, and the smaller group has the same weight as the larger one to enhance the influence of smaller groups. These two methods are also called *average linkage* clustering methods (26,1,25). The distance between two clusters is:

$$\mathcal{D}_{X,Y} = \frac{\sum \mathcal{D}_{x,y}}{n_X \cdot n_Y} \tag{4}$$

where X and Y are two clusters, x and y are objects from X and Y, $\mathcal{D}_{x,y}$ is the distance between x and y, and n_X and n_Y are the respective sizes of the clusters. In WPGMA, these two numbers are set to the higher number in both clusters.

Ward's Method. This method is based on the statistical minimization of clustering expansion (3). In the course of every step, the central point is calculated for any possible combination of two clusters. In addition, the sum of the squared distances of all elements in the clusters from their central points is computed. The two clusters that offer the smallest possible sum are used to formulate the new cluster. The notion of distance used here has no geometric nature.

General Algorithm

Before the grouping commences, objects following the chosen probabilistic guidelines are generated. In this article, objects are randomly selected and are drawn from the interval $[0, 1]^2$. Subsequently, the objects are compared to each other by computing their distances. The distance used in assessing the similarity between two clusters is called the *similarity coefficient.* This is not to be confused with *coefficient of correlations* as the latter are used to compare outcomes (i.e., hierarchical trees) of the clustering process. The way objects and clusters of objects coalesce together to form larger clusters varies with the approach used. Below, we outline a generic algorithm that is applicable to all clustering methods (initially, every cluster consists of exactly one object):

1. Create all possible cluster formations from the existing ones.
2. For each such candidate compute its corresponding similarity coefficient.
3. Find out the minimum of all similarity coefficients and then join the corresponding clusters.
4. If the number of clusters is not equal to one (i.e., not all clusters have coalesced into one entity), then go to step 1. Otherwise terminate.

Essentially, the algorithm consists of two phases: the first phase records the similarity coefficients. The second phase computes the minimum coefficient and then performs the clustering.

There is a case where ambiguity may arise when using average-based methods. For instance, let us suppose that when performing Step 1 (of the previous algorithmic skeleton), three successive clusters are to be joined. All these three clusters have the same minimum similarity value. When performing Step 2, the first two clusters are joined. However, when computing the similarity coefficient between this new cluster and the third cluster, the similarity coefficient value may now be different from the minimum value. The question at this stage is what the next step should be. There are essentially two options:

- continue by joining clusters using a recomputation of the similarity coefficient every time we find ourselves in Step 2, or
- join *all* those clusters that have the same similarity coefficient at once and do not recompute the similarity in Step 2.

In general, there is no evidence that one is better than the other (1). For our study, we selected the first alternative.

Statistical Distributions

As already mentioned, objects that participate in the clustering process are randomly selected from a designated area (i.e., $[0, 1] \times [0, 1]$). There are several random distributions; we chose three that closely model real world environments (3). Our aim is to examine whether clustering is dependent on the way objects are generated. We use three distributions for the creation of data, namely: uniform, piecewise (skewed), and finally Gaussian distribution. Next, we describe these statistical distributions in terms of distribution and density functions.

Uniform Distribution. The respective distribution function is the following: $\mathcal{F}(x) = x$. The density function of this distribution is $f(x) = \mathcal{F}'(x) = 1 \; \forall x$ such that $0 \leq x \leq 1$.

Piecewise (Skewed) Distribution. The respective distribution function is the following:

$$\mathscr{F}(x) = \begin{cases} 0.05 & \text{if } 0 \le x < 0.37 \\ 0.475 & \text{if } 0.37 \le x < 0.62 \\ 0.525 & \text{if } 0.62 \le x < 0.743 \\ 0.95 & \text{if } 0.743 \le x < 0.89 \\ 1 & \text{if } 0.89 \le x \le 1 \end{cases} \quad (5)$$

The density function of this distribution is: $f(x) = \mathscr{F}(b) - \mathscr{F}(a)/b - a$ $\forall x$ such that $a \le x < b$.

Guassian (Normal) Distribution. The respective distribution function is

$$\mathscr{F}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad (6)$$

This is a two-parameter (σ and μ) distribution, where μ is the mean of the distribution and σ^2 is the variance. The density function of the Gaussian Distribution is:

$$f(x) = \mathscr{F}'(x) = \frac{1}{\sqrt{2\pi}} \frac{\mu - x}{\sigma^3} e^{-(x-\mu)^2/2\sigma^2} \quad (7)$$

In producing samples for the Gaussian distribution, we choose $\mu = 0.5$ and $\sigma = 0.1$.

$$\mathscr{F}(x) = \begin{cases} 0.00132 & \text{if } 0.1 \le x < 0.2 \\ 0.02277 & \text{if } 0.2 \le x < 0.3 \\ 0.15867 & \text{if } 0.3 \le x < 0.4 \\ 0.49997 & \text{if } 0.4 \le x < 0.5 \\ 1 & \text{for } 0.0 \le x \le 1 \end{cases} \quad (8)$$

For values of x that are in the range [0.5, 1], the distribution is symmetric.

Two Examples

Here, we present examples of how data is clustered in order to illustrate how different clustering method work with the same set of data. Example 1 uses the Average while Example 2 demonstrates the work of the Ward method. The sample data set has 10 items and each item has an identification and coordinate values that help us calculate the distances. Data values are selected following the uniform distribution (see Table 1).

Table 1. Example of a Sample Data List (Ordered)

Id	X	Y
0	0.459162	0.341021
1	0.480827	0.865283
2	0.525673	0.180881
3	0.585444	0.802122
4	0.639835	0.405765
5	0.646148	0.600101
6	0.795807	0.841711
7	0.878851	0.586874
8	0.945476	0.105152
9	0.956880	0.168666

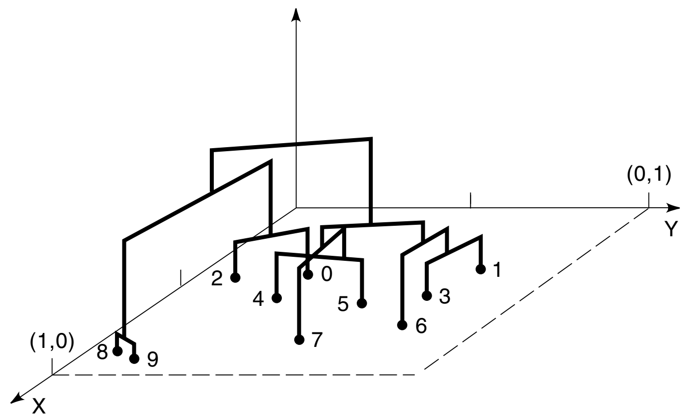

Figure 1. Clustering tree using Average.

Example 1. The steps described in this example give the progression of the algorithm while deploying the Arithmetic Average method with Unweighted Pair-Group. The dendrogram produced by this algorithm is shown in Fig. 1.

1. Join clusters {8} and {9} at distance 0.064530.
2. Join clusters {1} and {3} at distance 0.122205.
3. Join clusters {0} and {2} at distance 0.173403.
4. Join clusters {4} and {5} at distance 0.194439.
5. Join clusters {1, 3} and {6} at distance 0.264958.
6. Join clusters {4, 5} and {7} at distance 0.266480.
7. Join clusters {1, 3, 6} and {4, 5, 7} at distance 0.363847.
8. Join clusters {0, 2} and {8, 9} at distance 0.481293.
9. Join clusters {0, 2, 8, 9} and {1, 3, 6, 4, 5, 7} at distance 0.558245.

Example 2. The clustering of the two-dimensional sets of points using the Ward method is provided here. For each step we give the central point that results in the smallest squared sum of distances. The resulting dendrogram is shown in Fig. 2.

1. Clusters {8} and {9} maintain their central point at (0.951178, 0.136909) and join at distance 0.0020820397.
2. Clusters {1} and {3} have their central point at (0.533135, 0.833703) and joint at distance 0.0074670143.
3. Clusters {0} and {2} maintain their central point at (0.492418, 0.260951) and join at distance 0.0150342664.
4. Clusters {4} and {5} have their central point at (0.642992, 0.502933) and join at distance 0.0189031674.
5. Clusters {6} and {7} maintain their central point at (0.837329, 0.714292) and join at distance 0.0359191013.
6. Clusters {0, 2} and {4, 5} have their central point at (0.567704, 0.381942) and join at distance 0.1112963505.

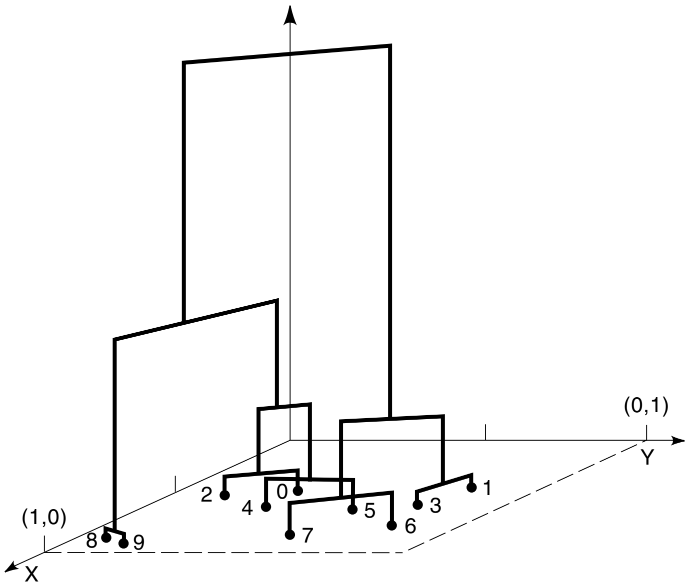

Figure 2. Clustering tree using Ward.

7. Clusters {1, 3} and {6, 7} maintain their central point at (0.685232, 0.773998) and join at distance 0.1217264622.
8. Clusters {0, 2, 4, 5} and {8, 9} have their central point at (0.695529, 0.300264) and join at distance 0.4144132554.
9. Clusters {0, 2, 4, 5, 8, 9} and {1, 3, 6, 7} maintain their central point at (0.691410, 0.489758) and join at distance 1.2178387810.

EXPERIMENTAL METHODOLOGY

The number of data items presented in this study ranges from 100 to 500; data are drawn from a two-dimensional (2-D) space and the values of the two coordinates range from 0 to 1 inclusive. In order to generate random sample data, the congruential linear algorithm described in (27,28) is used with the *seed* set to the computing system time.

Every conducted experiment goes through the following three steps:

1. Create the lists of objects.
2. Carry out the clustering process with the two different clustering methods (i.e., Average and Ward).
3. Calculate the coefficient of correlation for each clustering method.

For the purpose of obtaining a statistically representative clustering behavior, there is a need to repeat the same procedure a number of times. To achieve that goal, each experiment is repeated 100 times and the standard deviation of the coefficients of correlation is calculated. The least square approximation (LSA) is used to evaluate the acceptability of the approximation. If a correlation coefficient obtained using the LSA falls within the segment defined by the corresponding standard deviation, the approximation is considered acceptable.

The correlation coefficient is used as the main vehicle for comparing two trees obtained from lists of objects. The notion of distance used in the computation of the correlation coefficients could be realized in two ways: firstly, actual linear difference between any two objects could be used resulting in what is known as the Euclidean or linear difference. Secondly, the minimum number of edges in a tree that are required to join any two objects is used; this distance is termed the Edge difference. It is speculated that the latter way to compute the difference helps in a more “natural” implementation of a correlation. Once a distance type is chosen, we may proceed with the computation of the correlation coefficient. This is accomplished by first selecting a pair of identifiers (two objects) from a list (linearized tree) and calculating their distance and then by selecting the pair of identifiers from the second list (linearized tree) and computing their distance. We repeat the same process for all remaining pairs in the second list.

There are numerous families of correlation coefficients that could be examined. This is due to the fact that various parameters are involved in the process of evaluating clustering of objects in the two-dimensional space. More specifically, the clustering method is one parameter (i.e., Average or Ward); the method of computing the distances is another one (i.e., linear or edge); and finally, the distribution followed by the data objects (i.e., uniform, piecewise, and Gaussian) is a third parameter. In total, there are twelve (e.g., 2*2*3 = 12) possible ways to compute correlation coefficients for any two lists of objects. Also, the dimensional space added in this study may have a direct influence on the clustering. This determines what kind of data are to be compared and what their sizes are.

We have identified a number of cases to check the sensitivity of each clustering method with regard to the input data. For every type of coefficient of correlation previously mentioned, eleven types of situations (hence, eleven coefficients of correlation) have been isolated. All these types of situations are representative of a wide range of practical settings (16) and can help us understand the major factors that influence the choice of a clustering method (2,29,30).

We partition these settings into three major groups, represented by three templates or blocks of correlation coefficients.

First Block. The coefficients presented in this set examine the influence of context in how objects are finally clustered. In particular, the correlation coefficients are between:

1. Pairs of objects drawn from a set S and pairs of objects drawn from the first half of the same set S. The first half of S is used before the set is sorted.
2. Pairs of objects drawn from S and pairs of objects drawn from the second half of S. The second half of S is used before the set is sorted.
3. Pairs of objects drawn from the first half of S, say S_2, and pairs of objects drawn from the first half of another set S', say S'_2. The two sets are given ascending identifiers after being sorted. The first object of S_2 is given as identifier the number 1 and so is given the first object of S'_2. The second object of S_2 is given as identifier the number 2 and so is given the second object of S'_2 and so on.

4. Pairs of objects drawn from the second half of S, say S_2, and pairs of objects drawn from the second half of S', say S'_2. The two sets are given ascending identifiers after being sorted in the same was as the previous case.

Second Block. This set of coefficients determines the influence of the data size. Coefficients of correlation are drawn between:

5. Pairs of objects drawn from S and pairs of objects drawn from the union of a set X and S. The set X contains 10% new randomly generated objects.
6. Pairs of objects drawn as in case 5 but the set X contains 20% new randomly generated objects.
7. Pairs of objects drawn as in case 5 but the set X contains 30% new randomly generated objects.
8. Pairs of objects drawn as in case 5 but the set X now contains 40% new randomly generated objects.

Third Block. The purpose of this group of coefficients is to determine the relationship that may exist between two lists of two-dimensional objects derived using different distributions. More specifically, the coefficients of correlation are drawn between:

9. Pairs of objects drawn from S using the uniform distribution and pairs of objects drawn from S' using the piecewise distribution.
10. Pairs of objects drawn from S using the uniform distribution and pairs of objects drawn from S' using the Gaussian distribution.
11. Pairs of objects drawn from S using the Gaussian distribution and pairs of objects drawn from S' using the piecewise distribution.

In summary, all eleven types of coefficients of correlation are meant to analyze different settings in the course of our evaluation.

To ensure the statistical viability of the results, the average of one hundred coefficient of correlation and standard deviation values (of the same type) are computed. The least square approximation was then applied to obtain the following equation:

$$f(x) = ax + b \tag{9}$$

The criterion for a good approximation (or acceptability) is given by the inequality:

$$|y_i - f(x_i)| \leq \sigma(y_i) \quad \textit{for all } i \tag{10}$$

where y_i is the coefficient of correlation, f is the approximation function and σ is the standard deviation for y_i. If this inequality was satisfied, then f was a good approximation. The least square approximation, if acceptable, helps predict the behavior of clustering methods for points beyond the range considered in our experiments.

EXPERIMENTAL RESULTS

As stated earlier, the aim of this article is to conduct experiments to determine the stability of clustering methods and how they compare to each other. For the sake of readability, an abbreviated notation is used to indicate all possible cases. A similar notation has been used in our previous findings (16). For instance, to represent the input with the parameters Average, Uniform distribution, and Linear distance, the abbreviation AUL is used. (See Table 2.)

Table 2. List of Abbreviations

Term	Shorthand
Average	A
Ward	W
Uniform Distr.	U
Gaussian Distr.	G
Piecewise Distr.	P
Linear Distance	L
Edge Distance	E

The derived results are presented in figures and tables. The figures generally describe the different types of coefficients of correlation. The tables on the other hand describe the least square approximations of the coefficients of correlations.

Analysis of the Stability and Sensitivity of the Clustering Methods

We first look at the different clustering methods and analyze how stable and sensitive they are to the various parameters. More specifically, we are interested in knowing how sensitive each clustering method is to the changes of key parameter values.

Average: Results Interpretation. We look at the behavior of the three blocks of coefficients of correlation values as defined in the section on Experimental Methodology. We then provide an interpretation of the corresponding results.

First Block of Coefficients of Correlation. Figure 3 shows the four first coefficients of correlation corresponding to various alternate settings described by the block as the size of the participating lists ranges from one to five hundred objects. In addition, the corresponding standard deviations curves for all the experiments are shown as well. The difference between curves computing with either linear (L) or edge (E) distances is consistently small across all experiments. We also note that the values obtained using L are consistently larger than those resulting from the application of edge distance E. This is due to the fact that when L is used, the distance between the members of two clusters is the same for all members of the considered clusters. However, when E is used, this may not be true (e.g., tree that is not height balanced) since the distance is equal to the number of edges connecting two members belonging to different clusters. In the case of Fig. 4 (and the subsequent Fig. 8) the difference is attenuated due to the use of different distributions. When the values in L and E are compared against each other, the trend among the four coefficients of correlation is almost the same. This points to the fact that the distance type does not play a major role in the final clustering.

The absolute values maintained by the first and second types of correlation throughout the range of objects are larger

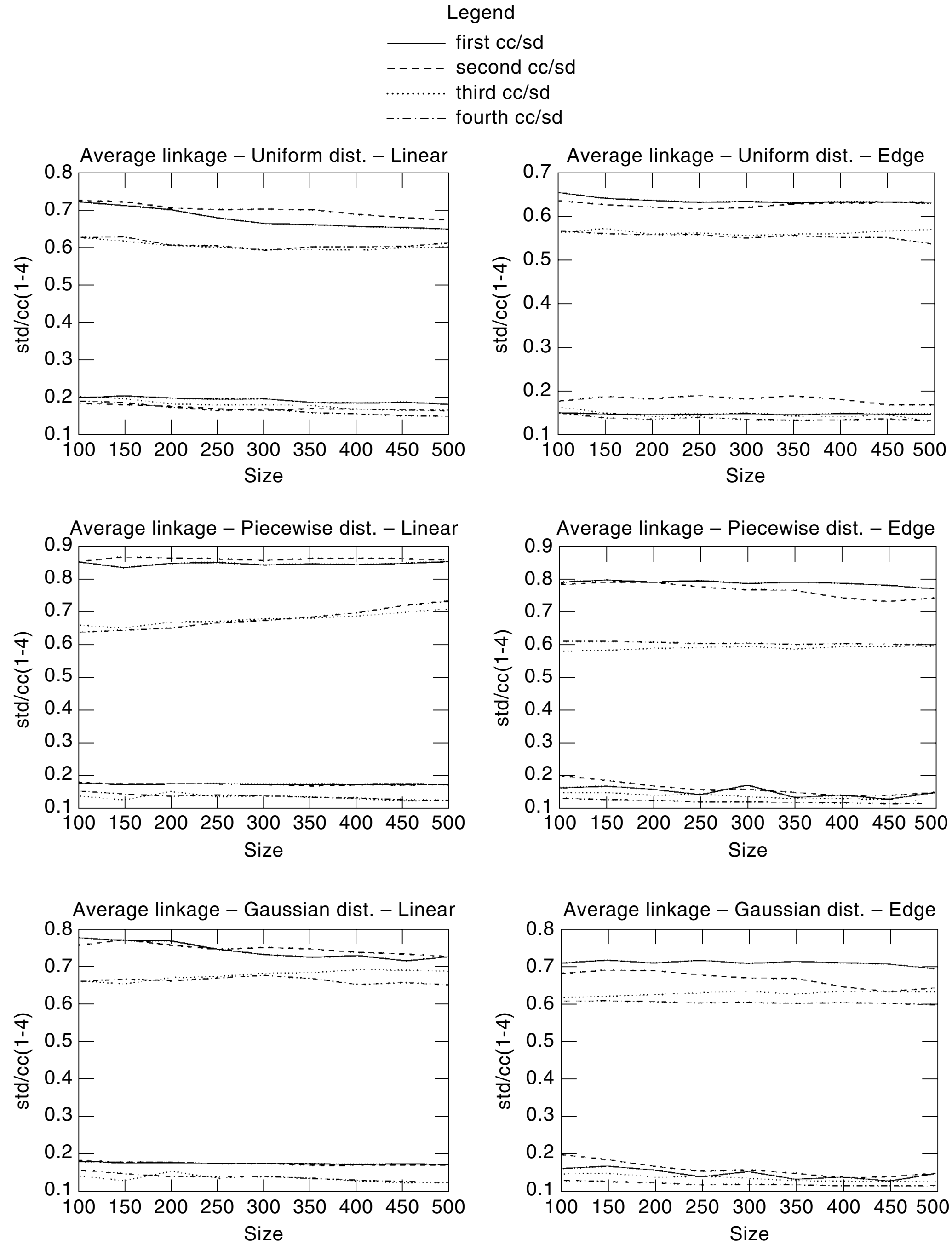

Figure 3. Average: first block of coefficient of correlation.

than their counterparts from the third and fourth types. This is attributed largely to the corresponding intrinsic semantics: the first and second types of correlations compare data objects drawn from the same initial set, whereas the third and fourth types of correlation associate data objects derived from different sets. This conforms to the expectation that objects from the first two correlations would be more closely related than data objects for the latter two. The standard deviation curves exhibit roughly the same behavior as the corresponding coefficient of correlation curves. This strongly suggests that the different types of correlation behave in a uniform and predictable fashion.

It is worthwhile noting that all the values for the correlation coefficients remain greater than 0.5 throughout all the graphs of Fig. 3. This fact implies that the data context does not seem to play an important role in the final data clustering. In a similar fashion, one can conclude that the data set size does not seem to have a substantial influence on the final

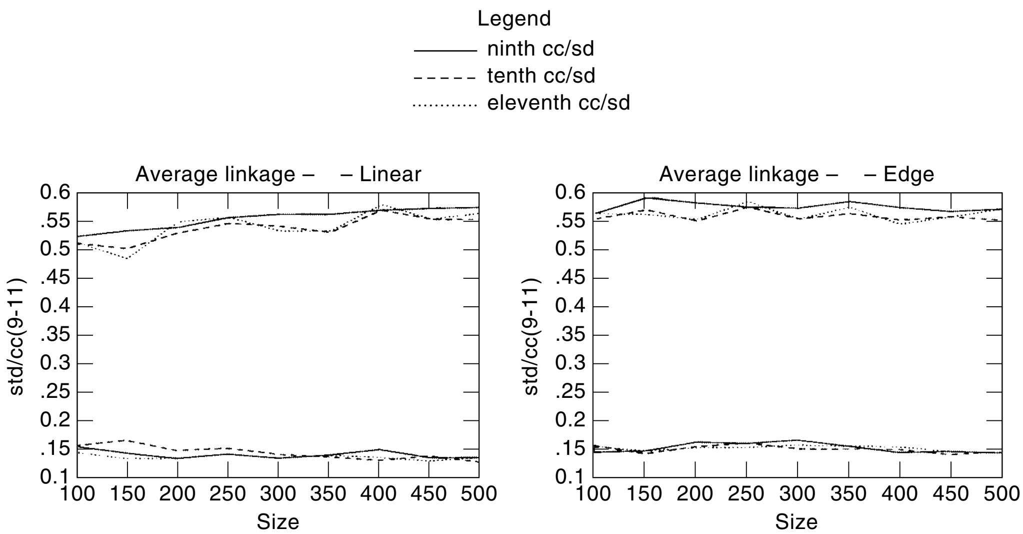

Figure 4. Average: third block of coefficient of correlation.

clustering. Note that the slope value is almost equal to zero. This is also confirmed by the uniform behavior across all the graphs of the standard deviation values above.

Second Block of Coefficients of Correlation. The experimental results discussed in this section examine the influence that the data size has on clustering. The produced graphs for the coefficients described by the second block are shown in Fig. 5. Both coefficient values and standard deviations are depicted as the number of objects participating in the experiments increases up to five hundred. The clustering method remains invariant (i.e., Average) while distance computations are performed with both linear and edge fashion using the three distributions.

There is no substantial difference between the curves computed using the linear (L) and edge (E) distances. This is indicative of the independence of the clustering from the type of distance used. The standard deviation values also exhibit the same behavior as one demonstrated by the corresponding coefficient of correlation values. This implies that the four types of correlation coefficients described by the second block maintain a uniform and predictable behavior despite the changes in the data sizes. The high values of the coefficients obtained suggest that the context sizes have little effect on how data is clustered. As in the previous case, the data size does not seem to influence the final clustering outcome very much as the slope (of the curves) is nearly equal to zero.

Third Block of Coefficients of Correlation. The subsequent three coefficients of correlation check the influence of the distribution for L and E. All other parameters are set the same for all pairs of objects in comparison. The curve representing the case for UP (Uniform and Piecewise distributions) in either L or E case demonstrates values lower than the corresponding values in the curves for both UG (Uniform and Gaussian distributions) and GP (Gaussian and Piecewise distributions). This can be explained by the problem of bootstrapping the random number generator. This is constant throughout most of the experiments conducted in this study. When the values in the cases of L and E are compared, no substantial difference is observed. This underlines the independence of the clustering from the two types of distances used. As the standard deviation values exhibit the same behavior as the corresponding coefficient of correlation values, uniform and predictable behavior of the different types of correlations is verified.

Since coefficient values converge around the value 0.5, this indicates that the distributions do not effect the clustering very much. The increase in the data size does not influence the final clustering outcome as the slope is nearly equal to zero. Therefore, the data set size does not have a substantial influence on the final clustering.

Ward: Results Interpretation. The results of the experiments using the Ward clustering method generally following the same type of pattern and behavior as the Average clustering method. Figure 6, Fig. 7, and Fig. 8 depict the first, second, and third blocks of coefficients of correlation. The interpretations that apply for the previous clustering method also apply for the Ward clustering methods as the resulting curves here follow a similar pattern of behavior. Indeed, the values for the coefficients of correlation and the standard deviations follow similar trends. In fact, there are few differences in the behavior of the Ward method as compared to the Average method.

Acceptability of the Least Square Approximation

Tables 3, 4, and 5 represent the least square approximations for all the curves shown in our study. The acceptability of an approximation depends on whether all the coefficients of correlation values fall within the interval delimited by the approximating function and the standard deviation. If this is the case, then we say that the approximation is *good.* Otherwise, we identify the number of points that do not fall within the boundaries and determine the quality of the function. Us-

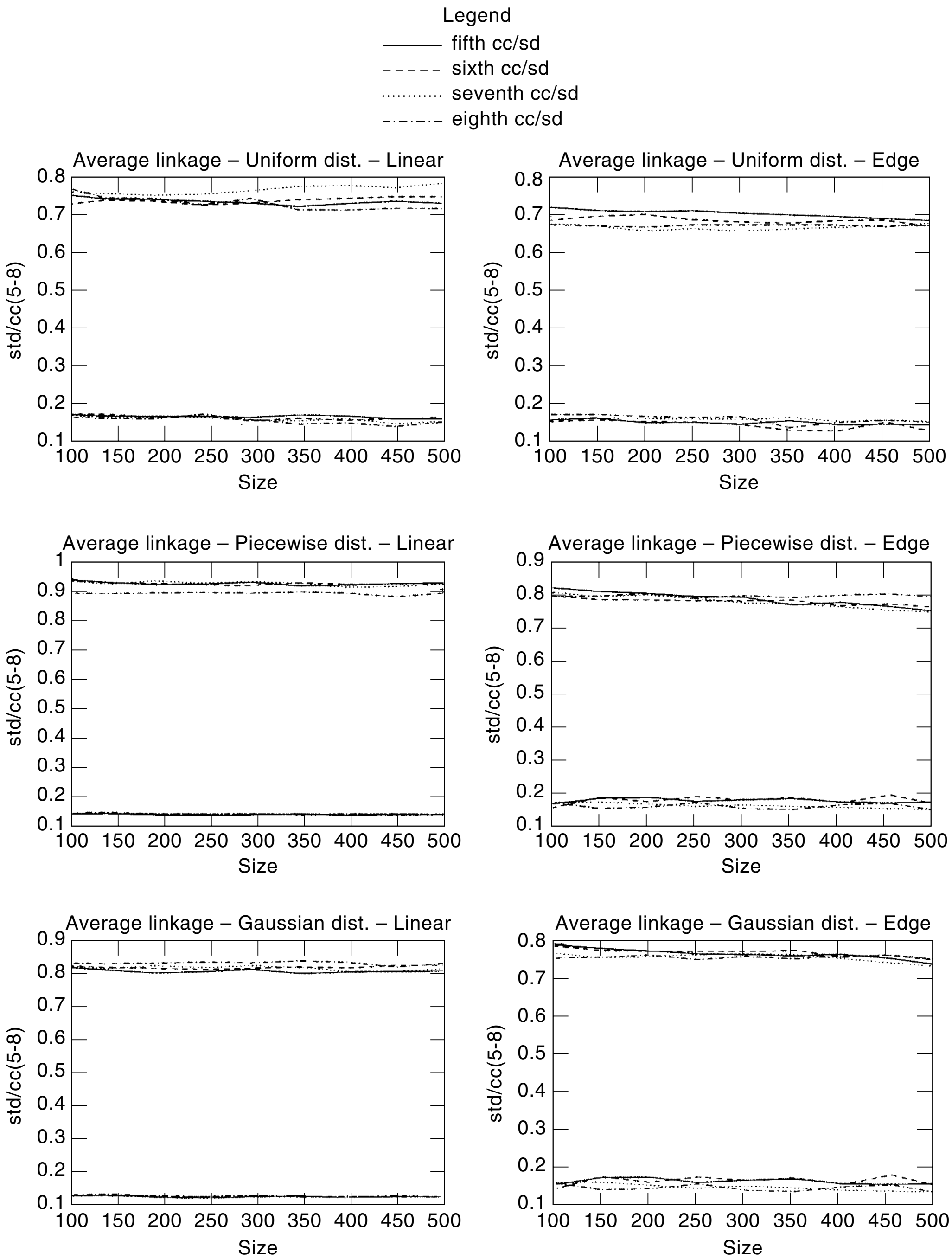

Figure 5. Average: second block of coefficient of correlation.

ing these functions enables us to predict the behavior of the clustering methods with higher data set sizes.

As all the tables show, the values of the slopes (derivatives) are all very small. This is indicative for the stability of all results. All approximations yield almost parallel lines to the x-axis. The acceptability test was run and all points passed the test satisfactorily. Therefore, all the approximations listed in the tables mentioned are good approximations.

Tabular Summary of Results for Average and Ward. Table 6 summarizes the results obtained using the Average and Ward clustering methods. Asymptotic values are used to provide a single value to represent the different clustering situations and for both clustering methods. The least square approximations are used as a tool for predicting and asymptotic values.

Block 1, Block 2, and Block 3 correspond to the first, second, and third block of correlation of coefficients described in a previous section. The summary points to a *high* level of similarity when asymptotic values are used when comparing the two methods. This should come as a surprise as the different parameters used do not seem to play any role in differentiating between the two methods. We provide a detailed comparative analysis in the next section.

Comparison of Results across Average and Ward Clustering Methods

In this section, we compare the different clustering methods against each other in light of the different parameters used in this study. These observations are drawn from the experiments and are shown in the presented figures and tables.

Context. The results show that across space dimensions, the context (i.e., where the objects are drawn) does not completely hide the sets. For instance, the first and second types of coefficients of correlation (as shown in all figures) are a little different from the third and fourth types of coefficient of correlation (as shown in all figures). The values clearly show that the context is *visible*.

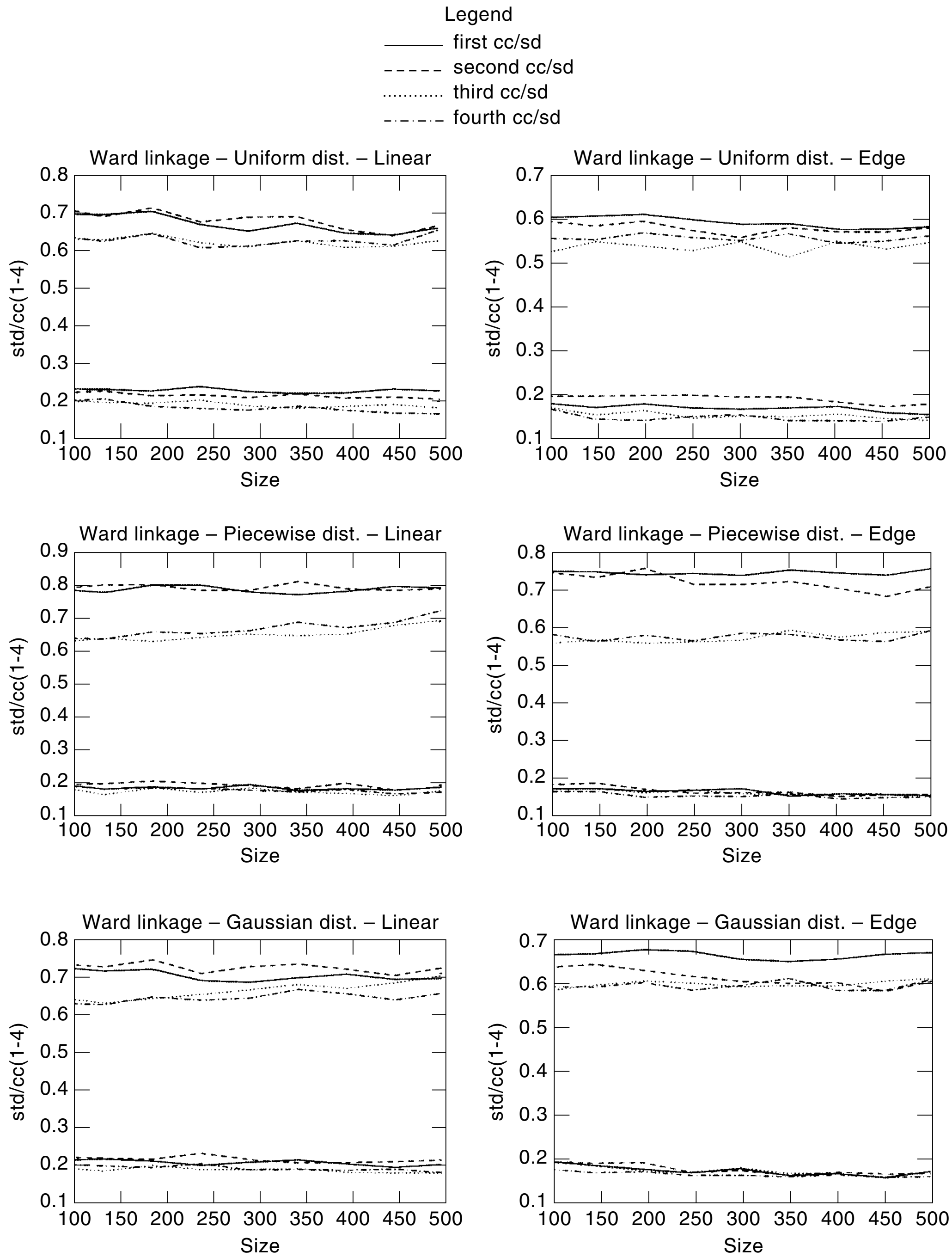

Figure 6. Ward: first block of coefficient of correlation.

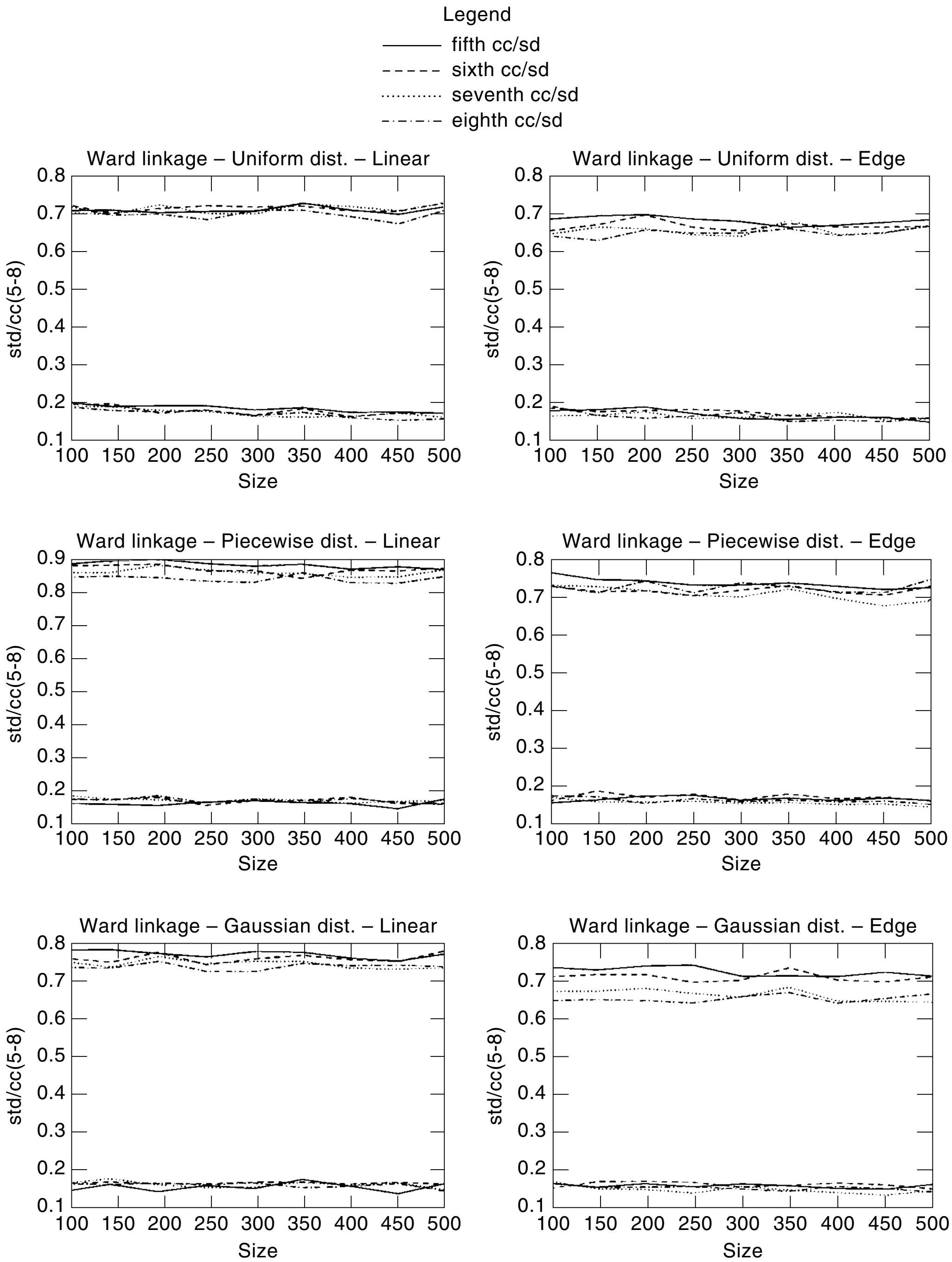

Figure 7. Ward: second block of coefficient of correlation.

The second block of coefficients of correlation for both clustering methods (fifth to eight coefficient of correlations, see Fig. 5 and Fig. 7), demonstrate that data size changes (perturbations) do not influence the data clustering because all coefficients of correlation values are high and somewhat close to 1.

Distribution. The results in all figures and Table 6 show that given the same distribution and type of distance, both clustering methods exhibit the same behavior and yield approximately the same values.

The results also show that the data distribution does not significantly affect the clustering techniques because the values obtained are very similar to each other (see Fig. 3, Fig. 5, and Fig. 6, Fig. 7, and Table 6). That is a relatively significant finding as the results strongly point to the independence of the distribution and the data clustering.

Stability. The results as shown in all figures also indicate that both clustering methods are equally stable. This finding comes as a surprise, as intuitively (because of the procedure

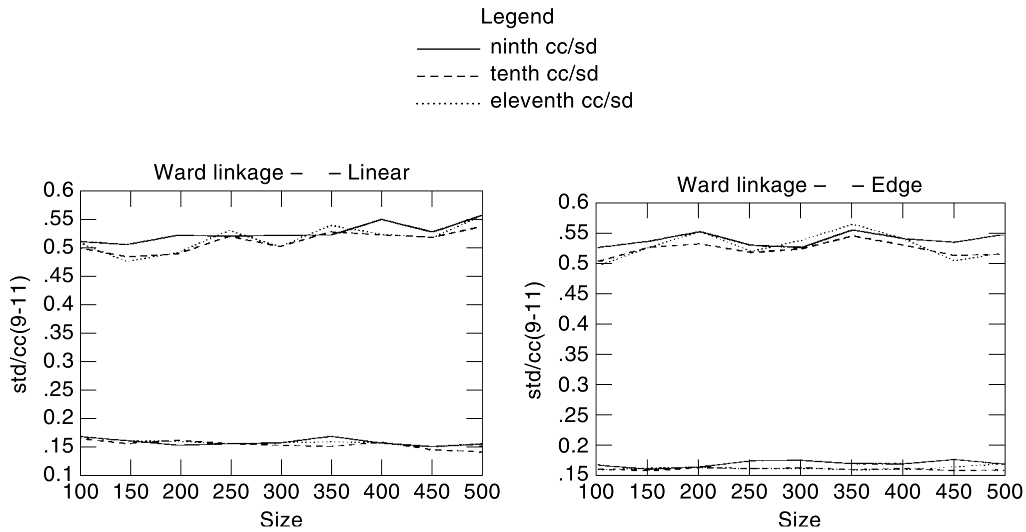

Figure 8. Ward: third block of coefficient of correlation.

Table 3. Function Approximation of the First Block of Coefficients of Correlation

	First Correlation	Second Correlation	Third Correlation	Fourth Correlation
AUL	0.000023 X + 0.72	0.00035 X + 0.73	0.00057 X + 0.62	0.00007 X + 0.63
AUE	0.00074 X + 0.67	0.00042 X + 0.65	0.00095 X + 0.59	0.00106 X + 0.58
APL	0.00061 X + 0.86	−0.00071 X + 0.88	−0.00074 X + 0.67	−0.0007 X + 0.66
APE	0.0000017 X + 0.81	0.000003 X + 0.80	0.0000096 X + 0.60	0.0000108 X + 0.62
AGL	0.00019 X + 0.78	0.00014 X + 0.77	−0.00084 X + 0.67	−0.00086 X + 0.67
AGE	0.0000054 X + 0.72	0.00059 X + 0.69	−0.000095 X + 0.63	0.000116 X + 0.61
WUL	−0.000009 X + 0.70	0.000001 X + 0.71	0.00043 X + 0.64	0.00051 X + 0.64
WUE	0.000063 X + 0.63	0.000044 X + 0.61	−0.0009 X + 0.56	0.00093 X + 0.58
WPL	−0.0000029 X + 0.80	−0.0004 X + 0.81	−0.00045 X + 0.64	−0.0000034 X + 0.65
WPE	−0.00012 X + 0.76	0.00023 X + 0.76	−0.0000074 X + 0.57	−0.0000082 X + 0.59
WGL	0.00022 X + 0.71	0.000004 X + 0.72	−0.0000055 X + 0.63	−0.000055 X + 0.62
WGE	0.0000057 X + 0.68	0.0007 X + 0.64	−0.00076 X + 0.60	−0.00086 X + 0.60

Table 4. Function Approximation of the Second Block of Coefficients of Correlation

	Fifth Correlation	Sixth Correlation	Seventh Correlation	Eighth Correlation
AUL	0.00028 X + 0.75	−0.0003 X + 0.74	−0.00019 X + 0.76	0.0004 X + 0.76
AUE	0.000059 X + 0.72	0.00073 X + 0.70	0.00065 X + 0.67	−0.00063 X + 0.67
APL	−0.00051 X + 0.93	−0.00052 X + 0.93	−0.00053 X + 0.93	−0.00025 X + 0.89
APE	0.0001 X + 0.81	0.000013 X + 0.78	0.00031 X + 0.79	0.000022 X + 0.79
AGL	0.0000041 X + 0.81	0.00032 X + 0.82	0.00033 X + 0.82	−0.0000023 X + 0.83
AGE	0.0000026 X + 0.78	0.00023 X + 0.77	0.00047 X + 0.76	0.00049 X + 0.75
WUL	−0.0000019 X + 0.71	−0.00023 X + 0.72	−0.0003 X + 0.71	−0.000038 X + 0.71
WUE	0.00044 X + 0.70	0.00044 X + 0.68	−0.00056 X + 0.66	−0.00059 X + 0.65
WPL	−0.0000055 X + 0.89	−0.000053 X + 0.88	−0.00048 X + 0.87	−0.000033 X + 0.84
WPE	0.00022 X + 0.76	0.00026 X + 0.72	0.000045 X + 0.73	0.00031 X + 0.73
WGL	0.0002 X + 0.78	0.00014 X + 0.76	0.0000027 X + 0.75	0.0000033 X + 0.74
WGE	0.0000032 X + 0.74	0.00036 X + 0.72	0.00068 X + 0.68	−0.00066 X + 0.66

Table 5. Function Approximation of the Third Block of Coefficients of Correlation

	Ninth Correlation	Tenth Correlation	Eleventh Correlation
AL	$-0.0000116\,X + 0.52$	$-0.00114\,X + 0.51$	$-0.0000121\,X + 0.51$
AE	$0.00092\,X + 0.58$	$-0.00096\,X + 0.56$	$-0.00094\,X + 0.56$
WL	$-0.00093\,X + 0.51$	$-0.00098\,X + 0.49$	$-0.00093\,X + 0.49$
WE	$-0.0000082\,X + 0.53$	$-0.00092\,X + 0.52$	$-0.00089\,X + 0.52$

in computing the distances), one expects the Average clustering method to show more stability than Ward.

Clustering Behavior. The third block of coefficients of correlation (see Fig. 4 and Fig. 8) across both clustering methods show that the two methods are little or not perturbed even in a noisy environment since there are not significant differences in results from Uniform and Piecewise, and Gaussian distributions. In addition, it is important to mention that the standard deviation small values (around 0.2) for all methods as shown in the figures seem to suggest a relatively high behavior stability. This important characteristic holds independently from any changes in *all* the parameters considered for this study.

Distance Used. The type of distance (linear or edge) as shown in all figures does not influence the clustering process as there are not significant differences between the coefficients of correlation obtained using either linear or edge distances.

These findings are in line with earlier findings (16) where one-dimensional data samples and fewer parameters were utilized. The results obtained here tend to indicate that *no* clustering technique is better than the other when data are drawn from a two-dimensional space. What this essentially means is that there is an inherent way for data objects to cluster, and independently from any technique used. The second important result this study seems to suggest is that the sole discriminator for selecting a clustering method should be based on its computatational attractiveness. This is a significant result as in the past there was no evidence that clustering methods exhibited similar patterns of behavior (1).

Table 6. Summary of Results

		Average	Ward
	Block 1		
L	U	0.65	0.65
	P	0.8	0.75
	G	0.7	0.7
E	U	0.6	0.6
	P	0.7	0.65
	G	0.65	0.65
	Block 2		
L	U	0.75	0.7
	P	0.9	0.85
	G	0.8	0.75
E	U	0.7	0.7
	P	0.75	0.75
	G	0.75	0.7
	Block 3		
L		0.55	0.55
E		0.55	0.55

SUMMARY

As clustering enjoys increased attention in data analysis of various computing fields such as data mining, access structures, and knowledge discovery, the study of the quality of various alternative methods becomes imperative. In this paper, we study the stability and behavior of two such clustering techniques, namely: the unweighted pair-group using arithmetic averages (termed Average) and Ward clustering. Data objects are drawn from a two-dimensional space following three different statistical distributions. In the course of our evaluation two types of distances are used to compare the resulting trees: the Euclidean and Edge distances. An exhaustive set of experiments is carried out in order to determine the various characteristics that the two methods offer. The three key results of this study are:

1. The Average and Ward clustering methods offer similar behavior and produce directly comparable results in a large number of diverse settings. We speculate that this similarity is attributed to the aggregate way distances are computed in order to determine similarity distances.
2. The two methods produce stable results.
3. The distributions of the two-dimensional data as well as the type of distances used in our exhaustive experiments do not affect the clustering techniques.

The outcomes presented here are a strong indication that clustering methods in the two-dimensional space do not seem to influence the outcome of the clustering process. Indeed, both clustering methods considered here exhibit a behavior that is almost constant regardless of the parameters used in comparing them. Future work includes examination of the stability of various clustering techniques in the three- and multidimensional data spaces and studying the effects that the various data-related and clustering parameters have in divisive methods.

ACKNOWLEDGMENTS

This work was partly funded by a QUT-NR grant number 160500 0015 and the center for Cooperative Information System (CIS) at QUT (for A. Bouguettaya) and the Center for Advanced Technology in Telecommunications (CATT) in Brooklyn, NY (for A. Delis).

BIBLIOGRAPHY

1. H. C. Romesburg, *Cluster Analysis for Researchers,* Malabar, FL: Kireger Publishing Company, 1990.
2. L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data, an Introduction to Cluster Analysis.* London: Wiley, 1990.

3. J. Zupan, *Clustering of Large Data Sets,* Letchworth, England: Research Studies Press, 1982.
4. F. Murtagh, A survey of recent advances in hierarchical clustering algorithms, *The Computer J.,* **26** (4): 354–359, 1983.
5. M. Tsangaris and J. F. Naughton, A Stochastic Approach for Clustering in Object Bases. In *Proc. Int. Conf. Management Data (SIGMOD),* 1991.
6. E. Ramussen, Clustering Algorithms in Information Retrieval, in W. B. Frakes R. Baeza-Yates (eds.), *Information Retrieval: Data Structures and Algorithms.* Englewood Cliffs, NJ: Prentice Hall, 1990.
7. D. Huntchens and V. Basili, System structure analysis: clustering with data bindings, *IEEE Trans. Softw. Eng.,* **11**: 749–757, 1985.
8. A. Delis and V. R. Basili, Data binding tool: A tool for measurement based on source reusability and design assessment, *Int. J. Softw. Eng. Knowl. Eng.,* **3** (3): 287–318, 1993.
9. A. K. Jain, J. Mao, and K. M. Mohiuddin, Artificial neural networks, *Computer,* **29** (3): 31–44, 1996.
10. C. T. Yu et al., Adaptive record clustering, *ACM Trans. Database Syst.,* **2** (10): 180–204, June 1985.
11. V. Benzaken and C. Delobel, Enhancing performance in a persistent object store: clustering strategies in O2. In *Proc. Conf. Principles Database Syst.,* 1990.
12. Jia-bing, R. Cheng, and A. R. Hurson, Effective Clustering of Complex Objects in Object-Oriented Databases. In *Proc. Int. Conf. Management Data (SIGMOD),* 1991.
13. W. J. McIver and R. King, Self-Adaptive, On-Line Reclustering of Complex Object Data. In *Proc. Int. Conf. Management Data (SIGMOD),* 1994.
14. V. Benzaken, An Evaluation Model for Clustering Strategies in the O2 Object-Oriented Database System. In *Proc. Int. Conf. Database Theory,* 1990.
15. J. Banerjee et al., Clustering a dag for cad databases, *IEEE Trans. Softw. Eng.,* **14**: 1684–1699, 1988.
16. A. Bouguettaya, On-line clustering, *IEEE Trans. Knowl. Data Eng.,* **8**: 1996.
17. S. Ceri and G. Pelagatti, *Distributed Databases: Principles and Systems,* New York, NY: McGraw-Hill, 1984.
18. G. Piatetsky-Shapiro and W. J. Frawley (Eds.), *Knowledge Discovery in Databases.* Menlo Park, CA: AAAI Press, 1991.
19. U. M. Fayyad et al., (Eds.), *Advances in Knowledge Discovery and Data Mining,* Menlo Park, CA: AAAI Press/MIT Press, 1996.
20. R. Ng and J. Han, Efficient and Effective Clustering Methods for Spatial Data Mining. In *Proc. 20th VLDB Conf.,* Santiago, Chile, 1994.
21. D. A. Bell et al., Clustering related tuples in databases, *The Computer J.,* **31** (3): 253–257, 1988.
22. C. Faloutsos et al., Efficient and effective querying by image content, *J. Intelligent Inf. Syst.,* **3** (3–4): July 1994.
23. B. Salzberg and V. J. Tsotras, A Comparison of Access Methods for Time-Evolving Data, Technical report, Northeastern University, 1994. NU-CCS-94-21, To Appear in *Computing Surveys.*
24. M. S. Alderfer and R. K. Blashfield, *Cluster Analysis,* Thousand Oaks, CA: Sage Publications, 1984.
25. B. Everitt, *Cluster Analaysis,* Yorkshire, England: Heinemann Educational Books, 1977.
26. G. N. Lance and W. T. Williams, A general theory for classification sorting strategy, *The Computer J.,* **9** (5): 373–386, 1967.
27. D. E. Knuth, *The Art of Computer Programming,* Reading, MA: Addison-Wesley, 1971.
28. W. H. Press, *Numerical Recipes in C: The Art of Scientific Programming,* 2nd ed., Cambridge University Press, 1992.
29. N. Jardine and R. Sibson, *Mathematical Taxonomy,* London: Wiley, 1971.
30. M. Tsangaris and J. F. Naughton, On the Performance of Object Clustering Techniques. In *Proc. Int. Conf. Management Data (SIGMOD),* 1992.

A. BOUGUETTAYA
Q. LE VIET
Queensland University of Technology
A. DELIS
Polytechnic University