LARS Contract Report 050279

Multidimensional Scaling for Clustering of Dissimilar Data Types

by **Nim-Yau Chu**
Paul E. Anuta

Prepared for
The National Science Foundation
Engineering Division
Washington, D. C.
Program Director: **Dr. Norman Caplan**

Grant No. ENG-7614400

Submitted by
Laboratory for Applications of Remote Sensing
Purdue University
West Lafayette, Indiana 47907

LARS CONTRACT REPORT 050279

Multidimensional Scaling for Clustering of Dissimilar Data Types

by

Nim-Yau Chu
Paul E. Anuta

ABSTRACT

The use of clustering algorithms for unsupervised classification of multi-dimensional data sets has come into wide usage in the remote sensing field. The data variables analysed generally come from the same general type of sensors and thus have similar dynamic ranges. When dissimilar data types such as encountered in geophysical remote sensing data sets, e.g., gravity, magnetics, gamma ray, dynamic ranges and distributions may be very different. The work reported here studies the problem of clustering algorithm performance relative to data range and normalization methods.

This research was sponsored by the National Science Foundation under Grant ENG-7614400.

Multidimensional Scaling for Clustering of Dissimilar Data Types

by

Nim-Yau Chu and Paul E. Anuta

Introduction

A problem in analyzing geophysical data is to find any structure hidden in the data set. A usual approach that most analysts will follow is to apply an unsupervised classification procedure or clustering to the data. Many cluster algorithms, such as those implemented in LARSYS, [6] utilize the "distances" between the data points. Naturally these distance-using methods are very sensitive to linear transformations of the data, since scaling often radically alters the interpoint distances. On the other hand, scaling may be used to correct some biased situations in which the data's numerical values of one feature are relatively high enough to dominate other features. Without any adjustment by scaling, clustering of all features simultaneously may be identical to clustering of the dominant feature alone.

The problem addressed here is to find a repeatable procedure which can determine a set of weights $w_1, w_2, \ldots, w_n$ of scaling for a data point $\underline{x}$ from a given n-feature data set. The scaled point $\underline{y}$ is given by

$$\underline{y} = \begin{bmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_n \end{bmatrix} \underline{x} \qquad (1)$$

Multidimensional scaling has been studied extensively in the discipline of behavioral sciences [1,2,3]. However, most of their methods and techniques may not be used directly for geophysical data owing to the nature of the data. Most data sets in behavioral sciences are qualitative descriptions taken from human subjects with personal variation. Therefore, scaling problems are oriented towards finding appropriate methods of assigning numerical values to qualitative descriptions and removal of subjective variation.

Geophysical data sets are usually relatively large; the sample size is often in the order of thousands. Such large sizes may justify easily the assumption that data tend to form clusters, and it makes better sense to talk about separability existing among clusters. We will see that this may lead to a method for determining the weights of scaling. For the purpose of controlled experimentation, we will use an artificial data set throughout this study.

Scaling: a tool to achieve equal importance among features

The effect of scaling on a clustering algorithm may be considered as modifying the implemented distance measure. Take, for example, the Euclidean distance used by ISODATA [4] clustering algorithm. The distance between two points $\underline{x}_1$ and $\underline{x}_2$ is given by:

$$d_x = \sum_{i=1}^{n} (x_1^i - x_2^i)^2 \qquad (2)$$

Now, if clustering is applied to the scaled points $\underline{y}_1$ and $\underline{y}_2$, then the distance is given by:

$$d_y = \sum_{i=1}^{n} (y_1^i - y_2^i)^2 = \sum_{i=1}^{n} w_i^2 (x_1^i - x_2^i)^2 = \sum_{i=1}^{n} W_k (x_1^i - x_2^i)^2, \quad (3)$$

where $W_k = w_i^2 \geq 0$.

The distance defined by (3) is known as modified Euclidean distance. It can be seen that if $W_k = 1$, (3) is the same as (2), so (2) may be considered a particular case of (3). What the Euclidean distance means is that the difference of $\underline{x}_1$ and $\underline{x}_2$ in all the features are contributed with equal importance to the distance. However by varying the weights in the modified Euclidean distance, the importance of each feature can be emphasized and de-emphasized, or even a feature may be excluded by setting its corresponding weight to zero.

It is natural for a clustering algorithm to adopt an unmodified distance measure rather than a modified distance measure because it has no prior knowledge of the relative importance of the features. However, as can be seen from (2), the actual importance depends on the numerical values of each feature relative to others. Unfortunately, the numerical values of a feature, often determined by the particular process by which that feature is measured, is not in general consistent with the concept of equal importance that has been built in to a clustering algorithm. Scaling is used to correct the elusive relative importance accidentally represented by the numerical values of the data.

Hence a method for determining weights of scaling consists of essentially a criterion to determine the relative importance among features and a procedure to carry out the criterion.

Clustering processors available at the Laboratory for Applications of Remote Sensing

There are four available processors, all of which are distance-using: (1) CLUSTER (in Larsys), (2) EIGENCLUSTER (in Larsysdv), (3) VARCLUSTER (in Larsysxp), and (4) ISOCLS (in EOD-Larsys).

The CLUSTER processor (Larsys User's Manual, Vol. II, p. CLU-17 to CLU-20) [11] uses the famous ISODATA algorithm [4, 5, 6]. Briefly the processor initializes a user-entered number of cluster centers along the diagonal of the parallelpiped whose boundaries are one standard deviation from the ensemble mean. Next it starts an iterative process of assigning data points to the nearest cluster centers by computing and comparing the Euclidean distances from the data points to the centers. After each iteration, it updates the positions of cluster centers and then repeats the iteration until a user-entered percentage of data points do not get reassigned.

The EIGENCLUSTER processor is almost identical to the CLUSTER, except that cluster centers are initialized with the help of principal component analysis.

The VARCLUSTER processor is a substantially modified version of EIGENCLUSTER to allow automatic reduction or increase of the number of clusters, if the processor sees appropriate. Still a user needs to specify the number of clusters for the purpose of initializing cluster centers. Combination of two clusters may occur if their transform divergence [6] is lower than a user-specified threshold value

Split of a cluster may occur if its variance is greater than a preprogrammed value, meanwhile this value is set to 3 for Landsat data.

The ISOCLS (The EOD-Larsys manual, p. 9-1 to 9-53) processor is a close relative of VARCLUSTER but different in the following aspects: (1) its distance measure is the L^1-norm; (2) it has no cluster center initialization. At the beginning of program execution, the ISOCLS takes the entire ensemble as one cluster, and splits it using the variance criterion; later it may combine using the transform divergence criterion. For more detailed comparisons, see [7].

There are some non-distance-using clustering algorithms. An example is the CLASSY [8,9]. CLASSY uses a prior knowledge to replace the distance measure. Another interesting clustering processor is the AMOEBA [10] which is a spatial clustering algorithm.

An example showing the effect of scaling

An artificial two-feature data set consisting of three well-separated clusters of Gaussian samples is chosen to demonstrate the effect of scaling. The numerical values of the data, as shown by the scatter plot in Fig. 1, range from 15.14 to 226.8 for the first feature (x-axis) and 3.455 to 6.51 for the second feature (y-axis). This particular choice of feature values is intended to create a situation in which a feature dominates over other features. Furthermore, two of the clusters are deliberately arranged in such a way that they are separable only in the less dominant feature but not

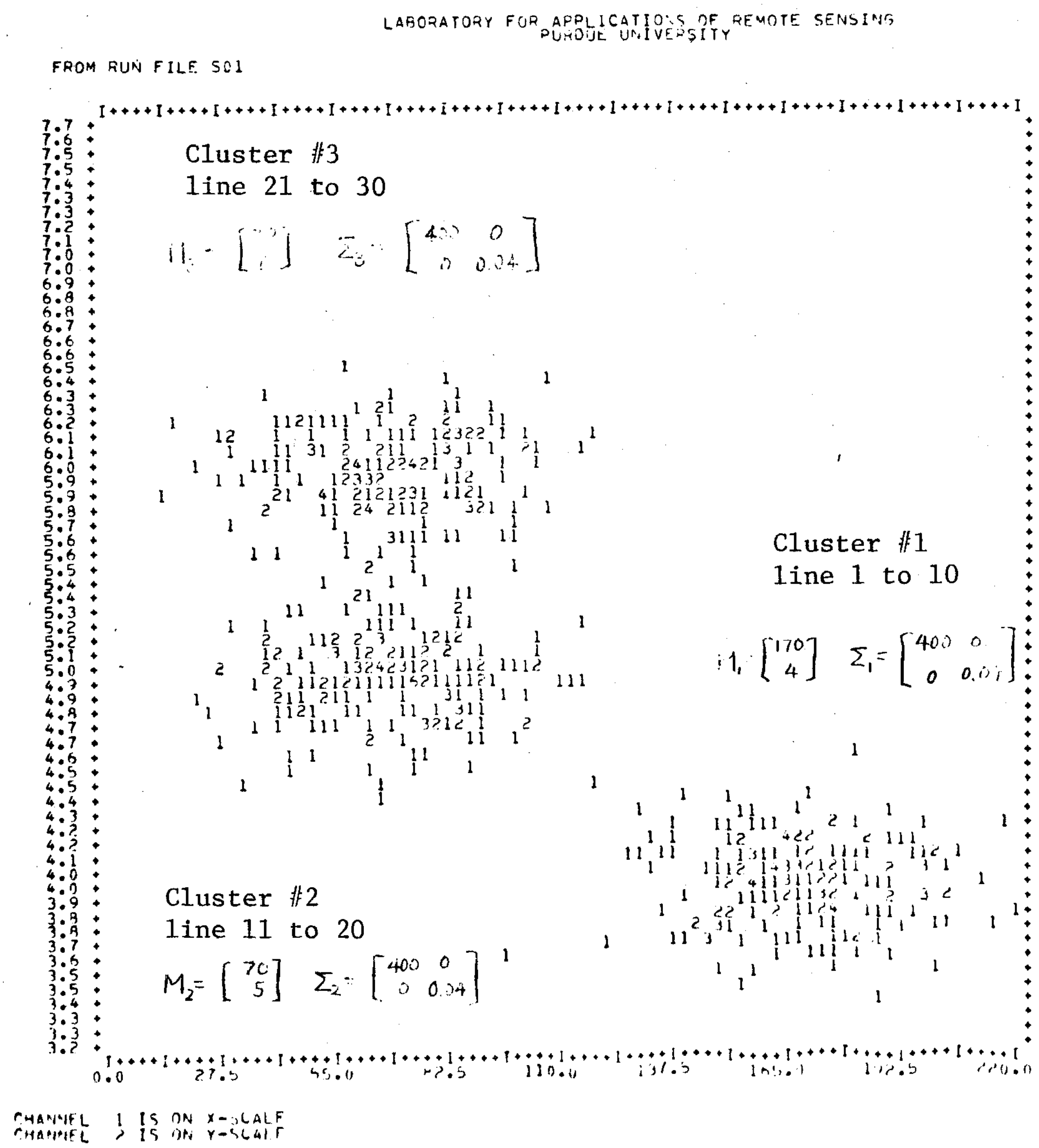
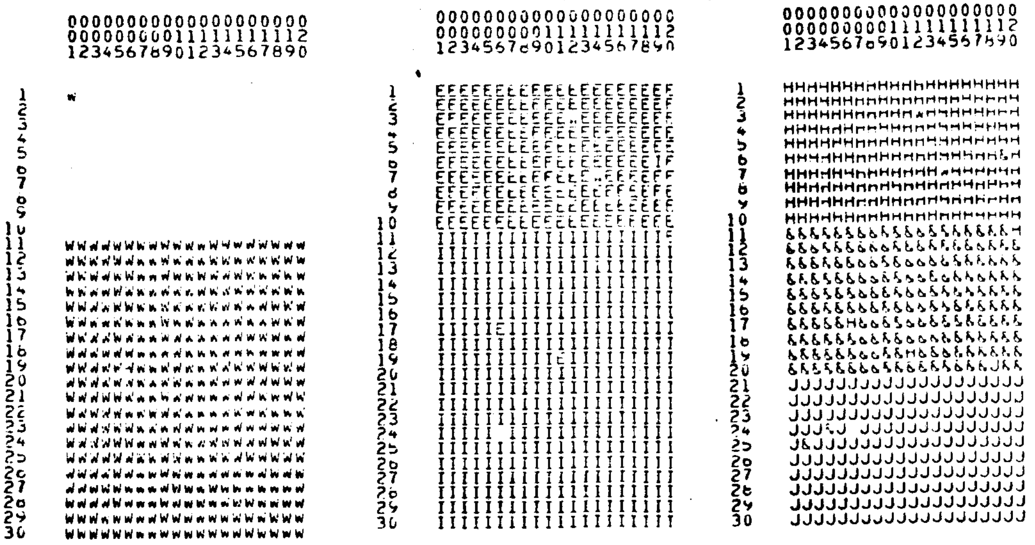


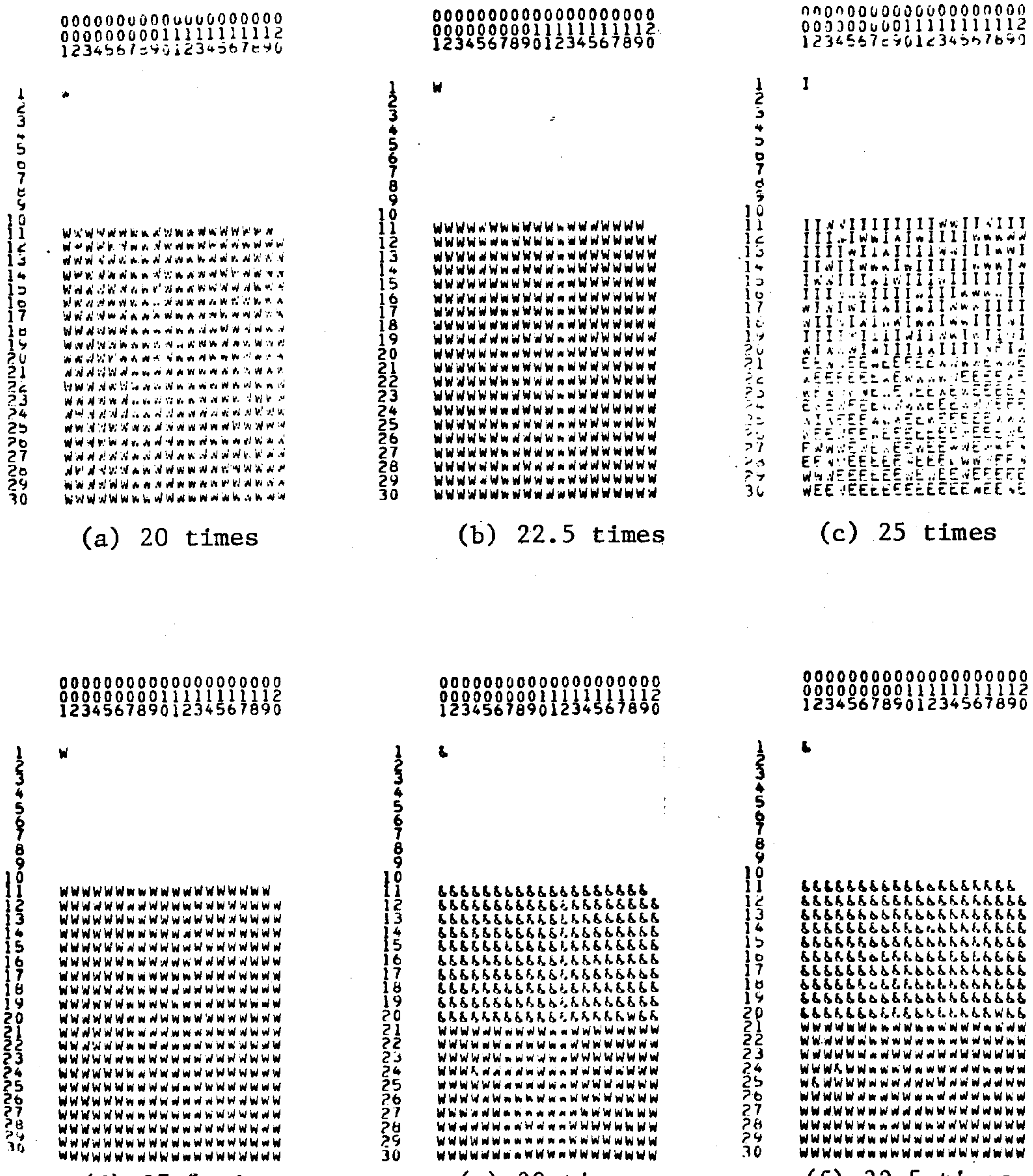
Fig. 1: Test data set I. This data set consists of three well-separated clusters of Gaussian samples, each cluster has 10x20 = 200 samples.

distinguishable in the dominant feature. Clustering using both features by a distance-using algorithm cannot reveal the above clusters easily. This is illustrated in Fig. 2(a) which is a cluster map created by VARCLUSTER processor using both features. It can be seen that the processor found two clusters only instead of three. This result is the same as the result obtained by clustering with the dominant feature alone, shown in Fig. 2(b). However, as shown in Fig. 2(c), the three clusters can be separated by clustering with the less dominant feature alone. The above exercise confirms that a hidden structure in some less dominant features may not be revealed in the presence of a dominant feature. Next we apply a sequence of scaling that alters the relative dominance or importance of the features, and then perform clustering on the scaled data using both features. In this sequence of tests, the data values of the less dominant feature are multiplied by 20, 22.5, 25, 27.5, 30 and 32.5, while the data values of the dominant feature remain unchanged. The gradual increase of the multiplier has the effect of increasing the importance of the second feature relative to the first feature. The cluster maps are shown in Fig. 3(a) through Fig. 3(f). It can be observed that the cluster processor continued to fail to distinguish the two close clusters for any multiplier up to 27.5, but suddenly from 30 onwards, the two clusters were distinguishable. Further experiments show that all clusters can be revealed for multipliers up to 1000. The finding is summarized in Fig. 4.



(a) (b) (c)

Fig. 2: Cluster maps produced by VARCLUSTER on the test data set I: (a) using both features, (b) using the first (dominant) feature, and (c) using the second (less dominant) feature. Note the symbols serves for identifying clusters only.



(a) 20 times

(b) 22.5 times

(c) 25 times

(d) 27.5 times

(e) 30 times

(f) 32.5 times

Fig. 3: Cluster maps produced by VARCLUSTER on the scaled data of test data set I. The second feature is multiplied by the number indicated below each map, while the first feature remains unchanged.

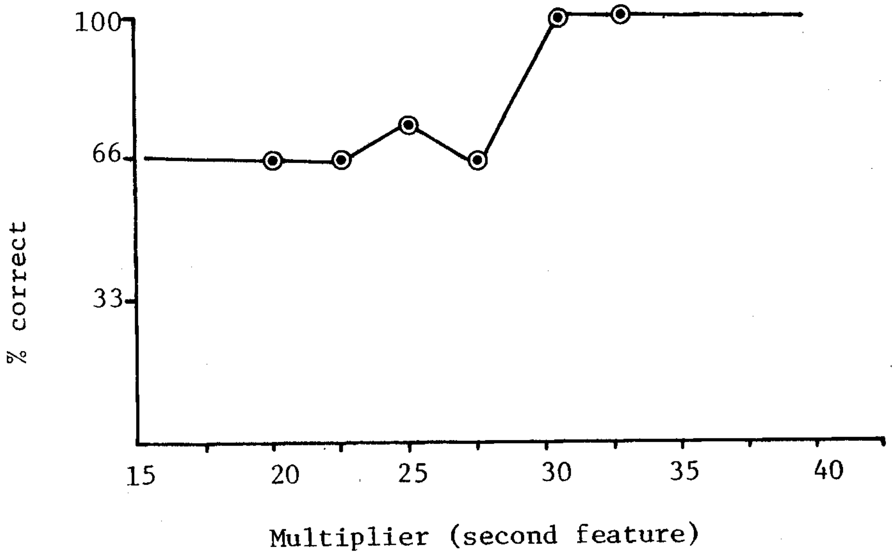


Fig. 4: Percentage of correctly clustered samples as function of the multiplier of the second feature for the test data set I.

Methods of determining the weights of scaling

There are three known methods of determining the weights of scaling, all of them are suitable for machine implementation.

(a) The dynamic range method

In this type of scaling, the data in each feature are linearly transformed (scaled) to a fixed dynamic range, for example, 0 to 255. All features, after scaling, have the same dynamic range. The idea of equal feature importance or dominance is now in the sense of equal dynamic range. Almost any type of data stored on magnetic tapes undergo this kind of scaling, therefore extra processing effort may not be necessary.

Procedure 1(a): Clustering with scaling determined by the dynamic range method (0-255)

(1) For each feature i, find from histogram the minimum value c_{i1} and the maximum value c_{i2}.

(2) Compute the scaled data using

$$Y_i = (x_i - c_{i1})*255/(c_{i2} - c_{i1})$$

(3) Cluster the scaled data

However if the geophysical data is available in Larsys MSS tape format and the ID record in calibration code 6 mode (i.e., the calibration values C1 is the minimum value and C2 is the maximum value), a much simpler procedure can be used.

Procedure 1(b): Clustering with scaling to fixed dynamic range (0-255) for data in LARS MSS tape format

(1) Use the following CHANNEL card
CHANNEL 6(1/0.,255./,2/0.,255./,..,n/0.,255./)
in *CLUSTER, *EIGENCLUSTER or *VARCLUSTER.

(b) The variance method

This method is similar to the dynamic range method, except that the scaled data now have unity variance. Since variance is a better measure of the spread of data than the dynamic range, thus it has a lower chance of mishandling the situation in which data appear to occupy a wide dynamic range but actually concentrate in a narrow range. Now the concept of equal feature importance is in the sense of equal variance.

Procedure 2(a): Clustering with scaling determined by the variance method

(1) For each feature i, find from statistics calculation the standard deviation of the feature, σ_i.

(2) Compute the scaled data:

$$y_i = x_i/\sigma_i$$

(3) Cluster the scaled data

Procedure 2(b): Clustering with scaling determined by the variance method for data in LARS MSS tape format

(1) For each feature i, find from the ID-record the minimum value c_{i1}, the maximum value c_{i2}, and from statistics calculation the standard deviation σ_i.

(2) Compute values $a_i = c_{i1}/\sigma_i$, $b_i = c_{i2}/\sigma_i$

(3) Use the following CHANNEL card

CHANNEL $6(1/a_1,b_1/, 2/a_2,b_2/,\ldots,n/a_n,b_n/)$

in *CLUSTER, *EIGENCLUSTER or *VARCLUSTER.

(C) The separability method

Since a geophysical data set is usually large and it seems more probable that the data form clusters. To be consistent with the primary objective of finding hidden clusters, the scaling should be such that it emphasizes a feature if clusters appear separable in that feature, and vice versa. Hence, the weights of the scaling are chosen in such a way that features having equal separability among possible clusters are equally important. The determination of weights involves single-feature clustering and cross-comparison of differences in the means of cluster pairs with comparable separability. Since differences in the means are compared, the separability is preferrably measured by normalized distance [6] defined as

$$d_{norm} = \frac{|\mu_1-\mu_2|}{\sigma_1+\sigma_2}$$

Scaling weights are the multipliers that are needed to bring the differences in the means to approximately the same value in all features for cluster pairs of comparable separability. For example, if in feature #1 a pair of clusters is 2 units apart and have a normalized distance of 3, and in feature #2, another pair of clusters (not necessarily the same pair in feature #1) is 100 units apart but have a normalized distance of 2.9. Since the normalized distances 3 and 2.9 are comparable to each other, therefore the scaling weights are determined as dividing 100 by 2, i.e., 50 for feature #1, and unity for feature #2.

Detection of separable clusters in a feature depends on solely single-feature clustering. This in turn demands a more powerful cluster processor. Since only separable clusters are of interest, the processor must be able to combine less separable clusters and delete unqualified clusters. Among the existing cluster processors available to users at LARS only the VARCLUSTER in Larsysxp and the ISOCLS in EOD-Larsys have this special capability.

After single-feature clustering, tables containing cluster pairs for each feature may be formed. Screening of validity of these pairs must be carried out to prevent misbehaving clusters from influencing the final decision of scaling weights. To guide screening, a more conventional separability measure such as transformed divergence may be used. Cluster pairs having separability below a threshold (i.e., not separable), e.g., 1300, should be discarded. Also clusters consisting of relatively few samples should be removed.

Procedure 3(a): Clustering with scaling determined by the separability method

(1) For each feature i, perform step (2) to (6)

(2) Do a single-channel clustering

(3) Delete clusters which (a) have only one sample; or (b) have a size less than a preprogrammed percentage of the average cluster.

(4) Form a table of cluster pairs. The first column is the absolute value of difference in the mean, next column is the normalized distance, and the third column is the transform divergence.

(5) (screening of the cluster pairs) Remove from the table the cluster pairs if (a) the value of transform divergence is less than a programmed threshold, say, 1300; or (b) the means of the cluster pair are equal.

(6) (Optional second screening of cluster pairs by clustering of the normalized distances as a data set) Cluster the normalized distances and select a subset which appears to form a group.

(7) For each pair in the table for the first feature, pick up those pairs in other features that are within a pre-programmed amount of its normalized distance.

(8) If not all features has a weight assigned, repeat step (7) with a less restrictive pre-programmed amount.

(9) (Determination of scaling weights) The weights are multipliers given by the ratios of their absolute values of difference in the mean

(10) Average the scaling weights

(11) Do linear transformation of the data

(12) Cluster the scaled data

In the above procedure, the steps (1) through (6) are for the preparation of cluster pair tables, one for each feature. The steps (7) through (10) are for the cross-comparison of normalized distances (which measures the separability) of cluster pairs from feature to feature. Such a comparison usually requires a complex decision rule and may be subject to instability. A simple but still effective alternative is to compute a mean-distance-to-normalized-distance ratio which simply indicates quantitatively a cluster pair's Euclidean distance per unit separability. The scaling weights are the ratios of these ratios.

Procedure 3(b): Weight determination by computing Euclidean distance per unit separability (steps to replace (7) to (10) in Procedure 3(a))

(7') For each feature i, do step (8) and (9)

(8') From the cluster pair table, compute and store in the fourth column of the table the mean-distance-to-normalized-distance ratio for all the cluster pairs.

(9') Compute the average ratio r_i from the fourth column of the table

(10') Find the smallest r_{min} from the ratios r_1, $r_2,\ldots,r_n$. Compute the scaling weights w_i by

$$w_i = r_i/r_{min}$$

Example of scaling weights determination

The above three methods for determining scaling weights were applied to the test data set I shown in Fig. 1. Information were extracted from the ID-record and statistics calculation are listed in Table 1(a). From these information, the weights can be determined easily by the dynamic range method and the variance method. These weights are shown in Table 1(b).

Table 1: (a) Information that can be used for determining scaling weights; (b) scaling weights.

(a)

	Min	Max	Standard Deviation
feature #1	15.14	226.8	51.17
feature #2	3.466	6.51	0.84

(b)

Method	Scaling Weights $w_1:w_2$	Comment
dynamic range	1:68	(226.8-15.14/(6.51-3.466)=68
variance	1:71	5.17/0.84 = 71

Since the separability method is more complex, the weight determination procedure is detailled presented as follows.

First, cluster pair tables are going to be derived from the single-feature clustering maps of Fig. 2(b) and 2(c). In the first feature, although there are three clusters found, actually there are two only because the third cluster is eliminated owing to too few samples. The cluster pair table, thus, has only one row, as shown in Table 2(a).

Table 2: Cluster pair tables

(a) Feature #1

Cluster ij	Dist. between the means	Normalized Distance	Transform Divergence	Mean distance/ normalized distance
1,2	100.3	2.56	1926	39.06
			average	39.06

(b) Feature #2

Cluster ij	Dist. between the means	Normalized Distance	Transform Divergence	Mean distance/ normalized distance
2,3	1	2.564	2000	0.390
3,4	1	2.5	2000	0.4
2,4	1	2.564	2000	0.39
			average	0.393

In the second feature, there are a total of 5 clusters found, but two of them, having too few samples, are discarded. Its cluster pair table is shown in Table 2(b).

Since transform divergences are high enough in both the tables, there is no further deletion of rows. Now the scaling weights are computed as the ratio of the mean distance-to-normalized-distance ratios, then the scaling weight is $w_1:w_2$ = 1:99.3(39.06/0.393).

It should be pointed out that if the second feature is multiplied by 100, then each cluster will have an equal variance Gaussian distribution. Regarding 1:100 as the best scaling weights, then we rate the separability method (1:99.3) as the best among the three methods. However, it was indicated previously that any multiplier larger than 30 is sufficient, so both the dynamic range method (1:68) and the variance method (1:71) provide adequate scaling, and furthermore they are much more simple procedure.

Data patterns that cause failure of the methods

There exist some data patterns that may make some scaling methods do not function properly or even fail.

A malfunction of the separability method occurs if a feature's histogram does not show (by multimodal) the possible clusters adequately. This is illustrated in Fig. 5(a) where the cluster #1 is four times larger than the other clusters, and it dominates the histogram of the second feature. Clustering of the second feature (Fig. 5(c)) only indicated presence of two clusters only. Although the separability method may still yield adequate scaling weights, but it did not function properly.

A second pattern that causes the failing of the variance method is shown in Fig. 6(a) where one of the clusters is located relatively far away from other clusters and yields a misleading large overall variance value. The standard deviations of the first and second features are found to be 51.17 and 3.1 respectively. The variance method will yield the scaling weights $w_1:w_2 = 1:16.5$ (51.17/3.1). The weights are not sufficiently large enough to bring about separation of the three clusters.

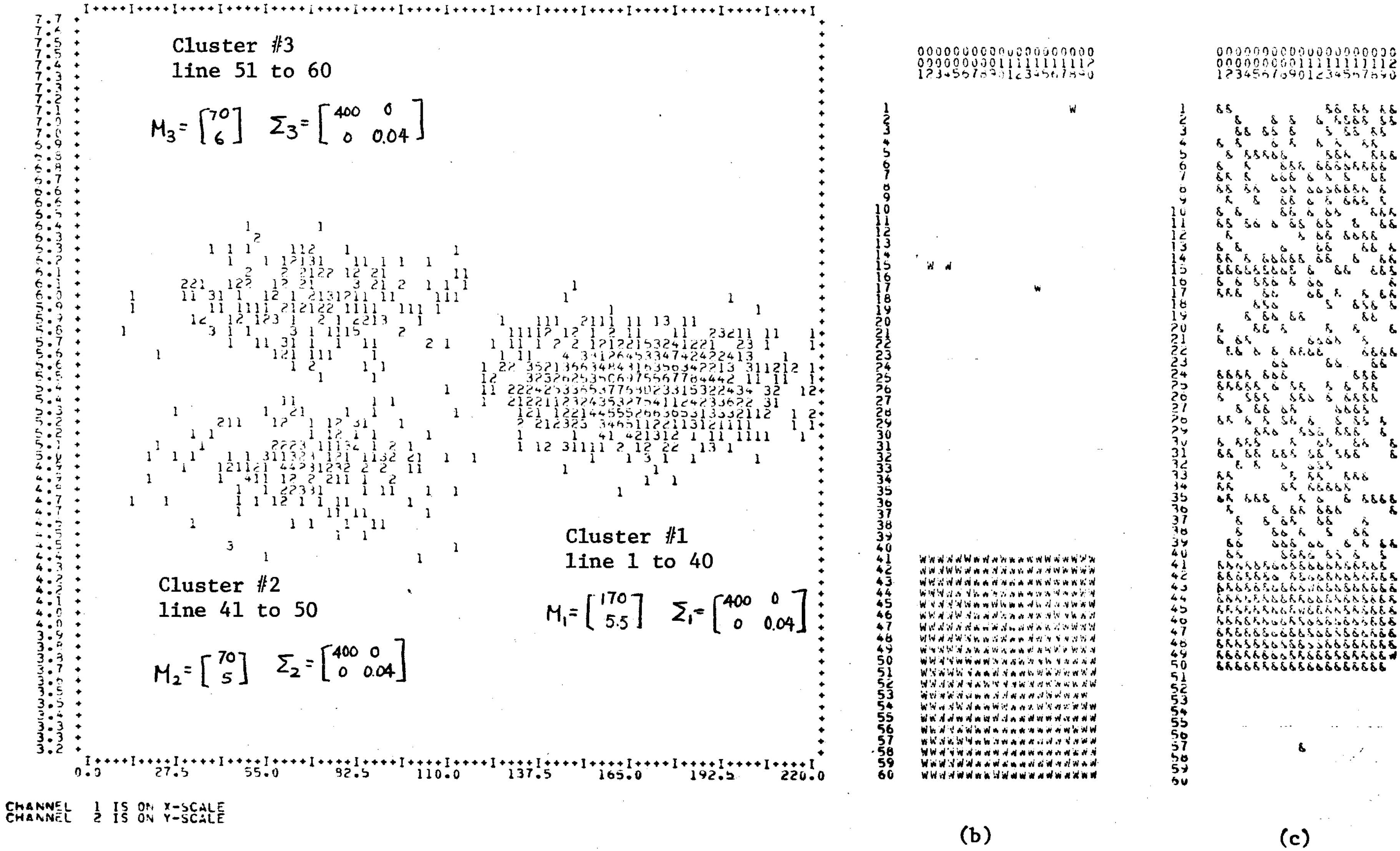


Fig. 5: Test data set II: (a) Scatter plot, (b) cluster map of first feature, and (c) cluster map of the second feature. The relative large sample size of cluster #1 handicaps the cluster processor, single-feature clustering of the second feature reveals only two clusters.

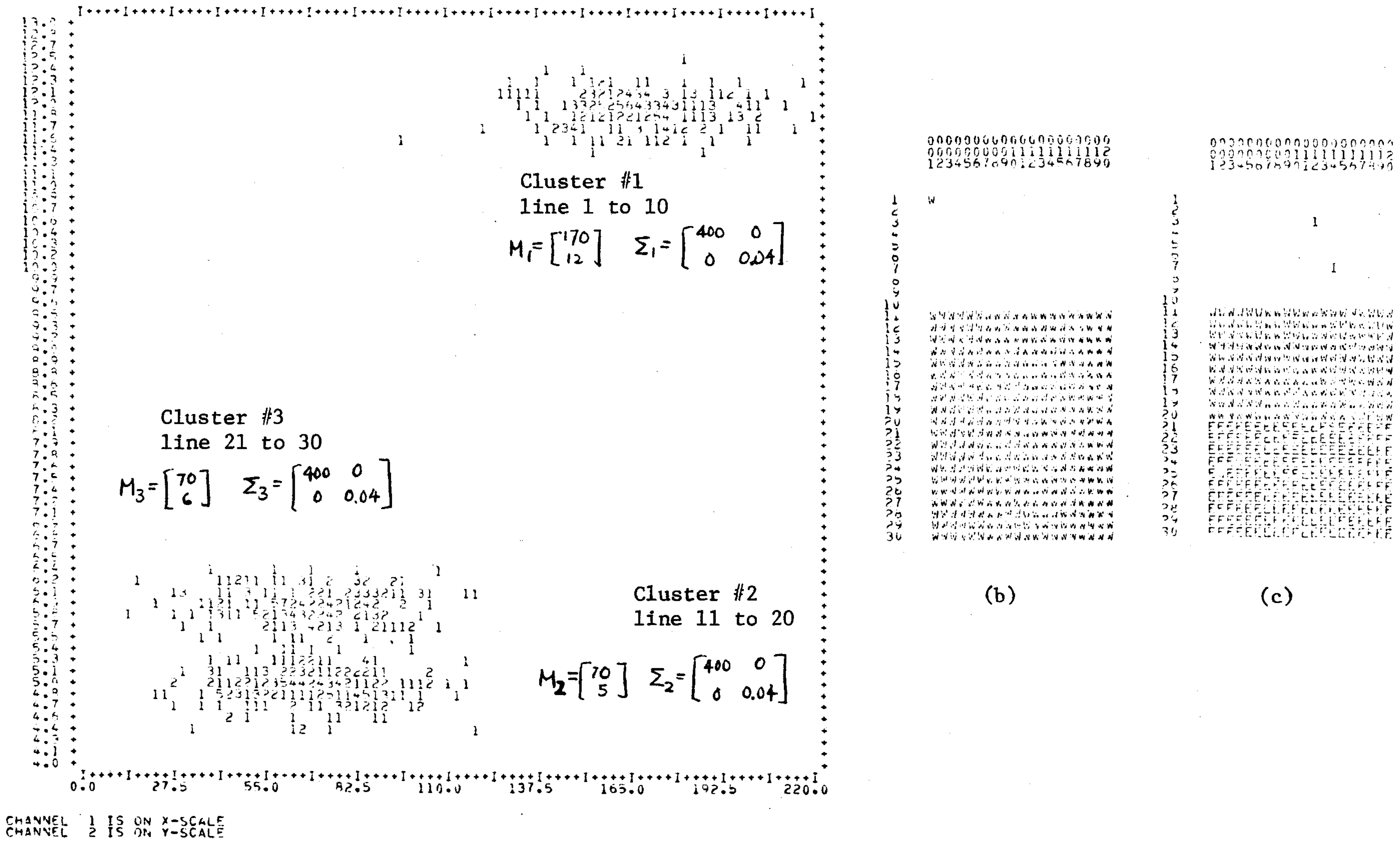


Fig. 6: Test data set III: (a) Scatter plot, (b) cluster map of the first feature, and (c) cluster map of the second feature. The cluster #1 is put at a distance far away from the other two clusters diliberately to make the variance of second feature large. The variance method of scaling may fail for this type of data pattern.

A third pattern that may cause failure of the separability method is shown in Fig. 7(a). The data, if investigated from histograms only, do not seem to contain four separate clusters, either viewed in the first feature only or viewed in the second feature only. This is an example of the situation where some clusters do not appear to be separable in single features but actually separable if all features are taken into consideration. Single feature clustering yields only two clusters. The scaling weights thereafter obtained are misleading.

Discussion of method failures

If it happens that the data, in some features, occupy a wide dynamic range but actually concentrate in a narrow range, the dynamic range method may yield insufficient scaling weights. However, this is an unlikely situation if data can be stored adequately on magnetic tapes where this type of scaling has been used implicitly. Also, this method as well as the variance method may fail in the situation where a feature has relatively few clusters relative to other features. Both methods tend to provide large weights for the feature having fewer clusters and thus over-emphasize that feature. The criterion of equal feature importance seems untrue.

It is observed that the separability method has a higher incidence of failure when using artificial data sets. The separability method depends on the availability of good single-feature clusters for weight determination. Too few clusters,

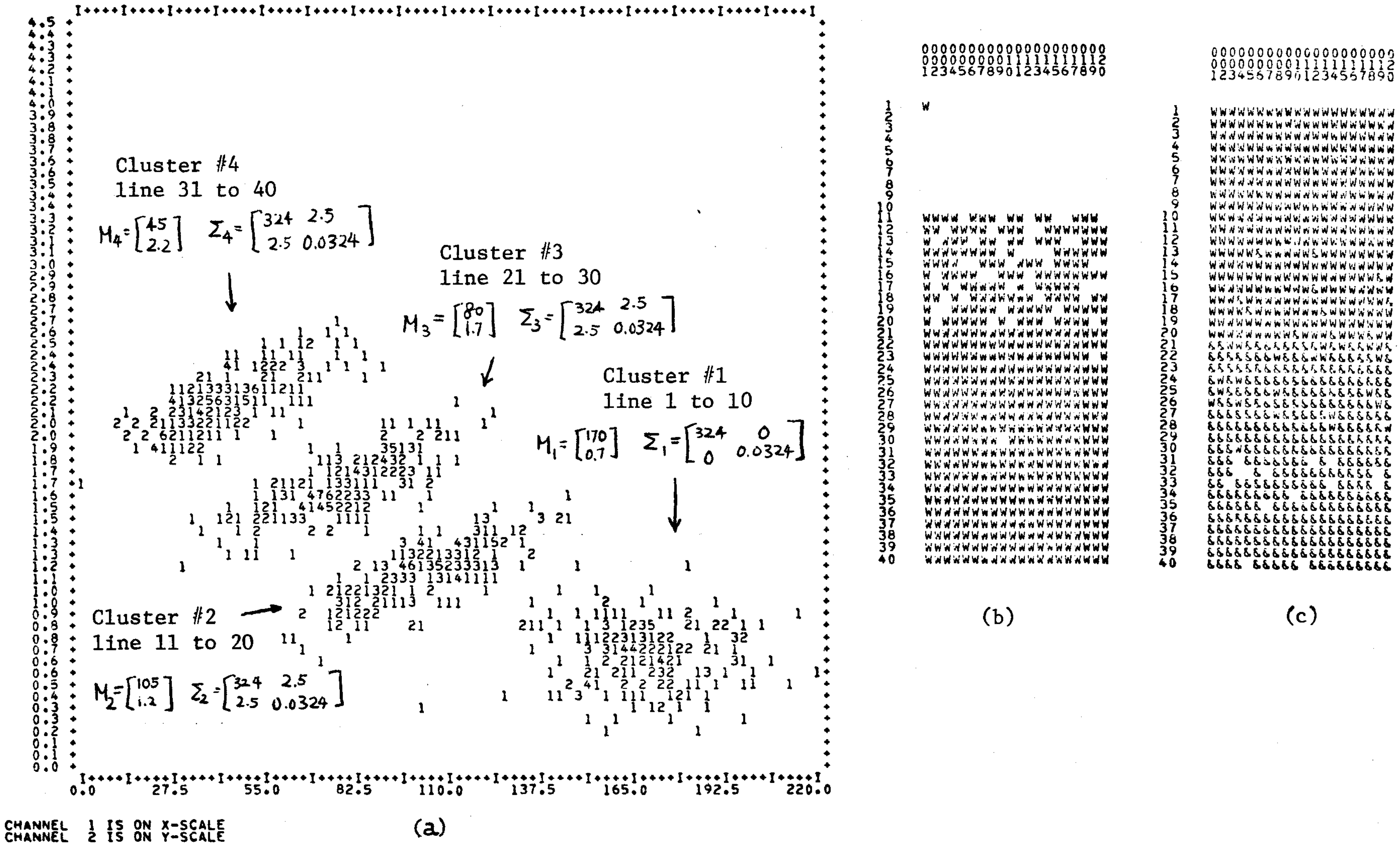


Fig. 7: Test data set IV: (a) Scatter plot, (b) cluster map of the first feature, and (c) cluster map of the second feature. Fewer than four separable cluster can be seen from each feature alone. The separability method of scaling may fail for this type of data pattern.

as in the case of artificial data sets (3 or 4), may fail to provide just two good clusters. However large data sets including geophysical data usually contain 5 to 15 separable clusters. With that many clusters, there is a better chance of having some good clusters after some misbehaving clusters are discarded.

Summary

It is conceived that any distance-using clustering algorithm has the built-in criterion of equal importance among all the features. However,due to the mechanism of clustering algorithms, this criterion is strongly influenced by the numerical values of the data, which unfortunately are not necessarily consistent with that criterion. The purpose of scaling is to transform the numerical values of the data to achieve consistency. Methods of determining the weights of scaling center on what sense the equal feature importance criterion is achieved. The dynamic range method is in the sense of equal dynamic range; the variance method is in the sense of equal variance; the separability method is in the sense of equal separability. Judging from the primary objective of finding hidden structure in a data set, the separability method appears to be the most superior in terms of criterion. As common to all criterion-using tools, no single criterion is successful for all types of data. There exist some data sets for which these methods may fail.

It is also conceived that all of the three methods will be successful for the type of data which contains relatively large (about ten) number of clusters, when the number of potential clusters are about the same in each feature and the sample size is large (thousands). Geophysical data belong to this type of data.

In many cases, which method to use depends on the cost. Obviously, the cost of using the dynamic range method is almost identical to the cost of clustering itself, the overhead cost is almost none. The variance method requires the knowledge of standard deviations, thus the overhead is the cost of performing statistics calculation. The separability method is the costliest of the three methods. Overhead includes the cost of those single-feature clusterings and examinations of the clusters.

References

[1] R. N. Sherpard, A. K. Romney and S. B. Nerlove, Multidimentional Scaling: Theory and Application in Behavioral Sciences, vol. I and II, Seminar Press, New York and London, 1972.

[2] J. B. Kruskal, "Linear transformation of multivariate data to reveal clustering," in Multidimensional Scaling: Theory and Application in the Behavorial Sciences, vol. 1, Theory. New York and London: Seminar Press, 1972.

[3] J. B. Kruskal, "Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new index of condensation," in Statistical Computation, R. C. Milton and J. A. Nelder, Ed. New York: Academic, 1969.

[4] Ball, G. H. and D. J. Hall, "ISODATA, A Novel Method of Data Analysis and Pattern Classification," Stanford Research Institute, Melno Park, California, 1965, pp. 1-16.

[5] Swain, P. H., "Pattern Recognition: A Basis for Remote Sensing Analysis", LARS Information Note 111572, Purdue University, 1972.

[6] P. H. Swain and S. M. Davis, Remote Sensing: The Quantitative Approach, McGraw-Hill, 1978, Chapter 3.

[7] Final Report on NASA contract NAS9-14016, Laboratory for Applications of Remote Sensing, 1975, pp. I-5 to I-14.

[8] McCray, B.: Program Documentation for Modifications to the CLASSY Program. LEC-10481, NASA/JSC (Houston), April 1977.

[9] Lennington, R. K.; and Malek, H.: The CLASSY Clustering

Algorithm - Description, Evaluation and Comparison with the Iterative Self-Organizing Clustering System (ISOCLS). LEC-11289, NASA/JSC (Houston), March 1978.

[10] J. Bryant, A Spatial Clustering Program: AMOEBA, Johnson Space Center, 1978.

[11] T. L. Phillips, Editor, "LARSYS Version 3 Users Manual," Laboratory for Applications of Remote Sensing Publication, Purdue University, June 1, 1973.