A Nonlinear Mapping for Data Structure Analysis

JOHN W. SAMMON, JR.

Abstract—An algorithm for the analysis of multivariate data is presented along with some experimental results. The algorithm is based upon a point mapping of \( N \) \( L \)-dimensional vectors from the \( L \)-space to a \( K \)-dimensional space such that the inherent data structure is approximately preserved.

Index Terms—Clustering, dimensionality reduction, mappings, nonparametric scaling, multivariate data analysis, pattern recognition, statistics.

INTRODUCTION

THE purpose of this paper is to describe the nonlinear mapping algorithm (NLM) which has been found to be highly effective in the analysis of multivariate data. The analysis problem is to detect and identify "structure" which may be present in a list of \( K \)-dimensional vectors. Here the word structure refers to geometric relationships among subsets of the data vectors in the \( L \)-space. Some examples of structure are hyperspherical and hyperellipsoidal clusters, and linear and certain nonlinear relationships among the vectors of some subset.

The algorithm is based upon a point mapping of the \( N \) \( L \)-dimensional vectors from the \( L \)-space to a lower-dimensional space such that the inherent structure of the data is approximately preserved under the mapping. The approximate structure preservation is maintained by fixing \( N \) points is the lower-dimensional space such that their interpoint distances approximate the corresponding interpoint distances in the \( L \)-space. We shall refer primarily to mappings in 2- and 3-dimensional spaces since the resultant data configuration can easily be evaluated by human observations in 3 or less dimensions.

THE NONLINEAR MAPPING

Suppose that we have \( N \) vectors in an \( L \)-space designated \( X_i, i = 1, \ldots, N \) and corresponding to these we define \( N \) vectors in a \( d \)-space (of \( d = 2 \) or \( 3 \) designated \( Y_i, i = 1, \ldots, N \) such that the distance between the vectors \( X_i \) and \( X_j \) in the \( L \)-space is preserved in the \( d \)-space with a corresponding small change in the distance between the vectors \( Y_i \) and \( Y_j \).

\[
Y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{id} \end{bmatrix}, \quad Y_j = \begin{bmatrix} y_{j1} \\ \vdots \\ y_{jd} \end{bmatrix}, \quad Y_k = \begin{bmatrix} y_{k1} \\ \vdots \\ y_{kd} \end{bmatrix},
\]

Next we compute all the \( d \)-space interpoint distances \( d_{ij} \) which are then used to define an error \( E \), which represents how well the present configuration of \( N \) points in the \( d \)-space fits the \( N \) points in the \( L \)-space, i.e.,

\[
E = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left[ d_{ij}^* - d_{ij} \right]^2.
\]

Let note that the error is a function of the \( d \times N \) variables \( y_{ij} \) \( p = 1, \ldots, N \) and \( y = 1, \ldots, d \). The next step in the NLM algorithm is to adjust the \( y_{ij} \) variables or equivalently change the \( d \)-space configuration so as to decrease the error. We use a steepest descent procedure to search for a minimum of the error (see Appendix I for further details).

Some Computer Results

We have exercised the nonlinear mapping algorithm on several data sets in order to test and evaluate the utility of the program in detecting and identifying structure in data. Some of the results obtained for several different artificially generated data sets are reported for the case where \( d = 2 \). We have also run the algorithm on many real data sets and have achieved highly satisfactory results; however, for demonstration purposes it is useful to work with artificially generated data in order that we can compare our results with the known data structure. The test data sets were as follows.

1) Straight Line Data: These data consisted of nine points distributed along a line in a 3-dimensional space.
2) The data points were spaced evenly along the line with an interpoint Euclidean distance of \( \sqrt{9} \) units. The initial 2-space configuration was chosen randomly.

For the purpose of this discussion it is convenient to think of the equipotential configuration as being selected randomly; however, in practice the initial configuration of the \( d \)-space is defined by projecting the \( L \)-dimensional data orthogonally onto a \( d \)-space spanned by the \( d \)-original coordinate axes.

One exception is data set 3 which is a classical data set. This data set was not artificially generated.
2) Circular Data: The data consisted of nine points, eight of which were spaced evenly (45° apart) along a circle of radius 2.5 units in a 2-dimensional space. The ninth point was placed at the center of the circle. The initial 2-space configuration was chosen randomly.

3) Iris Data: This data set is fairly well-known since it was used by Fisher [4] in several statistical experiments. The data were originally obtained by making four measurements on Iris flowers. These measurements were then used to classify three different species of Iris flowers. Fifty sample vectors were obtained from each of the three species. Thus, the data set consists of 150 points distributed in a 4-dimensional space.

4) Gaussian Data Distributed at the Vertices of a Simplex: These data consist of 75 points distributed in a 4-dimensional space. There are five spherical Gaussian distributions which have their respective mean vectors located at the vertices of a 4-dimensional simplex. The intervertex distance is 3/4 units and each covariance matrix is diagonal with a standard deviation along every coordinate of 0.2 unit. Fifteen points were generated from each of the five Gaussian distributions, making a total of 75 points.

5) Helix Data: This data set consisted of 30 points distributed along a 3-dimensional helix. The parametric equations for this helix are

\[ X = \cos Z \]
\[ Y = \sin Z \]
\[ Z = \sqrt{2} t / 2 \]

The points are distributed at one-unit intervals along the curve (i.e., \( t = 0, 1, 2, \ldots, 29 \)).

6) Nonlinear Data: This data set consisted of 29 points distributed evenly along a 3-dimensional curve. The parametric equations for this curve are

\[ X = \cos Z \]
\[ Y = \sin Z \]
\[ U = 0.5 \cos 2Z \]
\[ V = 0.5 \sin 2Z \]
\[ Z = \sqrt{2} t / 2 \]

The points were distributed at one-unit intervals along the curve (i.e., \( t = 0, 1, \ldots, 28 \)).

Figs. 1 through 9 display the results obtained using the nonlinear mapping algorithm. Convergence was essentially obtained for each case in twenty or less iterations using the gradient searching technique described in Appendix 1. As was expected, the data structure inherent in both data sets 1 and 2 was faithfully reproduced under the mapping (see Figs. 1 and 2). This was expected since the data sets are 1- and 2-dimensional, respectively, and therefore a mapping to a 2-space could be accomplished with zero error. In both cases the final error was 10⁻⁴. Observing the result of the Iris data mapping (Fig. 3), we can essentially detect these three species of Iris. The final error was 2 × 10⁻⁴ which is considered quite small.

The results obtained on the simplex data (data set 4) were quite interesting. The result of the mapping clearly showed the presence of five clusters (see Fig. 4). However, when we compare this result with the projection of the same data onto the 2-space defined by the two largest eigenvectors of the estimated data covariance matrix, we can only detect four clusters. Two of the clusters overlap completely in the 2-space which fits the data in the least squares sense (see Fig. 5). The resulting NLM error was 0.05. The same experiment has been conducted using Cauman data distributed at the vertices of higher-dimensional simplices. Figs. 6 and 7 show the NLM and principal eigenvector plots, respectively, for a 19-dimensional Gaussian simplex distribution. These experiments indicate that for some data sets, the NLM is superior to eigenvector projections for data structure analysis.

The results shown in Figs. 8 and 9 clearly indicate the "string structure" in data sets 5 and 6, respectively. The mapping error for the helix data was 6 × 10⁻⁴. The error for data set 6 was 1.6 × 10⁻⁴. The utility of any data analysis technique is somewhat more convincing when applied to "real" data as opposed to artificially generated data, presuming, of course, that the analysis results are correct. For this reason, the application of the NLM algorithm to an experiment document retrieval by content is reported here.

The experiment was conducted jointly by Rome Air Development Center (RADC) and the University of Colorado, involved the construction of a document classification space (referred to as the C-space) where every document in the library was represented as a 17-dimensional vector. The construction technique devised by Omori [8], [9] describes a mapping of 1125 president words and phrases into the C-space. Documents, or equivalently retrieval requests, were located in the space by computing the vector average of the corresponding key words or phrases which were contained in the document or the request. Retrieval was accomplished by rank-ordering the relevance of the library documents to a given request. The relevance measure was computed using the Euclidean metric between the document vectors and the request vector, the concept being that document vectors which are close in the C-space are related by content and therefore should be retrieved together.
Briefly, the C-space construction proceeded as follows. First the subject content covered by the 188 documents in the experimental library was subjectively partitioned into 23 technical fields (see Appendix II for a listing of these fields). Several experts representing each field rated the relevance of each of the 1125 words or phrases to its field, using a scale from 0 to 8. The rating by the experts within each field were then averaged to obtain a word-by-field relevance matrix designated X; the i,jth element of X represents the relevance of word or phrase i to field j. (It is convenient to think of the 1125 words or phrases as being represented by vectors in a 23-dimensional space spanned by the 23 coordinate fields.) Next, a 23x23 field correlation matrix C was computed, where the i,jth element represented the correlation between the i-th and j-th fields. C was then factored using the minimum residual method and rotated to a Varimax criterion. Seventeen orthogonal factors were then selected to define the 17-dimensional C-space. All 1125 words and phrase vectors were mapped into the 17-dimensional C-space using a simple nonlinear formula which tended to emphasize large coordinate
projections and minimize small coordinate projections. Finally, the 188 documents were located in the C-space by algebraically averaging the word or phrase vectors corresponding to the word or phrases which appeared in the documents. In order to evaluate the C-space as a potential method of document indexing, several individuals were asked to generate English queries (see Appendix III for the percent queries used here) which were then keypunched, automatically scanned for key word or phrase content, and finally mapped into the C-space. Each requester was then asked to identify those documents of the entire 188 which he felt were most relevant to his query. The C-space was then evaluated by examining the rank ordering of the retrieved documents to compare them to the list of relevant documents specified by the requester. The results of this evaluation can be found in Osorio [8]. The nonlinear mapping algorithm was used to evaluate the “structure” of the documents in the C-space. Specifically, we were interested in how the documents considered relevant to a particular request were clustered, and further, how these clusters were interrelated to each other and to the entire library. To accomplish
this analysis, all 188 17-dimensional vectors were used as the input data to the NLM. The numerals 1 through 5 were used in the resulting 3-dimensional mapping to designate the documents labeled relevant to queries 1 through 5. In addition, the symbol D was used to designate the remaining library documents. It is important to note that the NLM algorithm did not utilize the numeric query designations in computing the mapping. Only at the time of plotting the final 2-space configuration of the 188 points were the numeric and symbolic designators used to distinguish the data. The error in achieving the NLM shown in Fig. 10 was 0.662, which was considered to be acceptable for adequate 2-space representation.

The following facts were obtained upon investigation of the NLM result.

1) The documents considered relevant to a given request were clustered, lending evidence to support the hypothesis that related documents have C-space vectors which are close.

2) There does not appear to be any natural C-space structure relating subsets of documents. Instead, the documents tend to be uniformly distributed throughout the space.

3) Clusters 2 and 3 tend to overlap, yet they are well-separated from clusters 4 and 5. This can easily be accounted for since requests 2 and 3 are both concerned with the common subject of statistical data analysis, whereas 4 and 5 involve completely different subjects. In general, the intercluster relationships seem consistent with their respective subject relationships.

In summary, we have found the NLM algorithm to be of considerable value in aiding us in our understanding of the C-space as well as other document spaces. Presently we are planning to incorporate a similar mapping technique in an on-line document retrieval system in order to improve the retrieval via geometric means.

The experimental system will operate as follows. The on-line user would examine the 20 highest-ranked documents by retrieving and reading their abstracts. He would then indicate those he considered relevant. Next, a scatter diagram similar to Fig. 10 would be presented upon the CRT display. For each of the 38 documents it would be indicated by an I or an R, depending upon its relevance. In addition, the original query vector will be displayed as a Q. After examining the relative position of the documents in the mapping, the user would select (using a light pen) one or more relevant documents to be used to generate a new query vector(s). The concept is that the query vector can be moved to highly relevant regions of the document space by interacting at a display console with a geometric representation of the space.

RELATIONSHIP OF NLM TO OTHER STRUCTURE ANALYSIS ALGORITHMS

A mapping algorithm which bears a relationship to the NLM algorithm is one developed by Shepard [11] and later improved by Kruskal [5], [6]. Briefly, the Shepard-Kruskal algorithm seeks to find a configuration of points in a 2-space such that the resultant interpoint distances preserve a monotonic relationship to a given set of interelement similarities (or dissimilarities). Specifically, they wish to analyze a set of interelement similarities (or dissimilarities) given by $s_{ij}, i = 1, \ldots, N; j = 1, \ldots, N$. Suppose these similarities are ordered in increasing magnitude, such that $s_{12} \leq s_{13} \leq \cdots \leq s_{N-1,N}$.

The Kruskal-Shepard algorithm seeks to find a set of $N$ 2-dimensional vectors $y_i, i = 1, \ldots, N$, such that the order of the interpoint distances $d_{ij} = \text{dist}(y_i, y_j)$ deviates as little as possible from the monotonic ordering of the corresponding similarities. Although the mathematical formulations are similar, the underlying mapping criteria are quite different.

Ball [1] has compiled an excellent survey of clustering and agglomeration algorithms which are useful in solving the "structure analysis" problem. However, it has been our experience in using clustering techniques that these algorithms suffer to some extent from the following four deficiencies.

1) When using a particular algorithm, the resulting cluster configuration is highly dependent upon a set of control parameters which must be fixed by the user. Some examples of such parameters are:
   a) the similarity measure;
   b) various similarity thresholds;
   c) number of iterations required;
   d) thresholds which control the increase or reduction of the number of clusters;
   e) the minimum number of vectors required to define a cluster.
When choosing the control parameters for complex data, the user must either have a good deal of a priori knowledge regarding the "structure" of the data or be prepared to apply the algorithm many times for different values of the control parameters. This second alternative is, of course, tedious.

1. Most of the existing clustering algorithms are particularly sensitive to hyperspherical and hyperellipsoidal clusters and are inefficient in detecting more complex relationships in the data.

2. Perhaps the most serious deficiency involving present-day clustering algorithms is that these do not give really good ways for evaluating a resultant cluster configuration.

3. When two clusters are close, the vectors between them fail to form a bridge and cause spurious mergers [17].

We feel that the nonlinear mapping is a highly promising structure analysis algorithm since it suffers little from the listed clustering deficiencies. Consider the following facts concerning the algorithm:

1. The routine does not depend upon any control parameters that would require a priori knowledge about the data. Specifically, the user must set the number of mappings and the convergence constant \( M(F) \) in Appendix B.

2. It is highly efficient in identifying hyperspherical, hyperellipsoidal, and other complex data structures.

3. The resulting mapping (scatter diagram) is easily evaluated by the researcher, thereby taking advantage of the human ability to detect and identify data structures.

4. The problem concerning extraneous data and spurious mergers is not present since humans easily eliminate spurious data points by making global evaluations (machines have difficulty performing this function).

5. The algorithm is simple and efficient.

LIMITATIONS AND EXTENSIONS

There are, of course, limitations to every algorithm and the nonlinear mapping is no exception. There exist no limitations which we are presently investigating. The first has to do with the reliability of the scatter diagram in displaying extremely complex high-dimensional structure. It is conceivable that the minimum mapping error is too large \((2.00 \pm 0.1)\) and the 2-dimensional scatter plot fails to portray the true structure. However, we feel that for data structures composed of superpositions of hyperspherical and hyperellipsoidal clusters, the nonlinear mapping algorithm will, in general, display adequate representations of the true data "structure." The second limitation of the nonlinear mapping algorithm is related to the number of vectors that it can handle. Since we must compute and store the incidence matrix, which consists of \(N(N-1)/2\) elements, we are limited at present to \(N \leq 250\) vectors. In those cases where \(N > 250\), we suggest using a data compression technique to reduce the data set to less than 250 vectors. Specifically, we propose to use the \(k\)-data set [2] clustering algorithm to perform data compression. This is actually a natural function of clustering since we replace several vectors with a typical representative vector (i.e., the cluster center). Our previous objections to present-day clustering algorithms do not apply here since we are not concerned with fitting the data with 250 cluster centers. We are specifically not using the clustering algorithm to detect structure.

We have used the NLI to analyze multivariate data from two- or more classes for the purpose of determining how well the classes can be discriminated from one another. In these cases, it is recommended that the dimensionality be reduced to the smallest number of variables which preserves discrimination. In many problems certain measurements provide little discriminatory information; yet, if these measurements are included, the NLI will attempt to "fit" interpolating distances along these "noisy" directions as well as along discriminating directions. In truly high-dimensional problems, the resulting mapping may show considerable overlap between classes and still a high degree of discrimination may be possible. This phenomena occurred when analyzing a 4-class, 24-dimensional data set. The resulting NLI (the final error was 0.5, which was considered high) showed considerable overlap among less than 250 from three of the classes; yet, using a piecewise linear discrimination technique (based upon the use of a Fisher's linear discriminant between all pairs of classes), 94 percent correct classification was achieved. In this case, the NLI did not give an incorrect result since the...
IEEE TRANSACTIONS ON COMPUTERS, MAY 1960

\[ \frac{\delta \mathcal{E}}{\delta \gamma_i} = \frac{-2 \sum_{j \neq i} \frac{d_{ij}^* - d_{ij}}{d_{ij}^*} \left( y_{ij} - y_{ij}^\star \right)}{c} \]

and

\[ \frac{\delta \mathcal{E}}{\delta \eta_{ji}} = \frac{1}{c} \sum_{j \neq i} \frac{d_{ij}^* - d_{ij}}{d_{ij}^*} \frac{\delta y_{ij}}{\delta \eta_{ji}} \]

\[ \left( (d_{ij}^* - d_{ij}) \frac{1}{d_{ij}} (1 + \frac{d_{ij}^* - d_{ij}}{d_{ij}}) \right) \]

In our program we take precautions to prevent any two points in the d-space from becoming identical. This prevents the partials from "blowing up."

**APPENDIX II**

**Classification Space Files**

1. Adaptive Systems
2. Analog Computers
3. Applied Mathematics
4. Automata Theory
5. Computer Components and Circuits
6. Computer Memories
7. Computer Software
8. Display Consoles
9. Human Factors
10. Information Retrieval
11. Information Theory
12. Input-Output Equipment
13. Language Translation
14. Linear Algebra
15. Multivariate Statistical Analysis
16. Nonnumeric Data Processing
17. Numerical Analysis
18. Pattern Recognition
19. Probability and Statistics
20. Programming Languages
21. Stochastic Processes
22. System Design and Evaluation

**APPENDIX III**

**REQUESTS**

**Request 1:** What is known about the statistical distributions of words or concepts in English text? What impact does this knowledge or lack of knowledge have on the effectiveness of standard statistical methods in information retrieval problems? Are nonparametric methods more applicable?

**Request 2:** I am interested in techniques for data analysis. In particular, I wish information on "cluster-seeking" techniques as opposed to those of factorial analysis and discriminant analysis. "Cluster-seeking" techniques may be classified as follows: probabilistic techniques,
Mathematical Analysis of Ferrite Core Memory Arrays

WILLIAM T. WEEKS

Abstract—A mathematical model for simulating pulse propagation in a ferrite core memory array is described. Although specifically tailored to analyze 3-dimensional arrays, the model is sufficiently general to give a satisfactory analysis of pulse propagation, waveform distortion, and noise generation in a wide variety of memory configurations. The model treats the memory as a generalized, multiplied, multiconductor transmission line system. Under this assumption, the transmission line parameters are calculated from lumped parameters, thus leaving only a small number of parameters to be supplied empirically. Following a discussion of the equations which define the model and the methods by which they are solved, a sample array calculation is given to illustrate the kind of information that can be obtained from this model.

Johns Terms—Arrays, computer, ferrite core, memories, pulse propagation, transmission line system.

Manuscript received October 23, 1968; revised February 10, 1969.

The author is with IHP Corporation, Components Division, Hingham, N. Y. 12041.