Visualization of Data Structures and Machine Learning of Rules

Yoh-Han Pao

Case Western Reserve University
Cleveland, OH 44106
and
AI WARE, Inc.
Beachwood, OH 44122

Abstract

Intelligent task performing machines need to be able to benefit from experience and continue to improve its own task performance capabilities progressively. Towards that end, a machine capable of learning needs to be able to abstract bodies of complex high-dimensional data into manageable groupings and be able to discern and articulate relationships between such data items. This discussion describes how 2D depictions of multivariate data can support such Machine Learning activities. A new dimension-reduction procedure is described schematically. It seems to generate mappings which have useful 'topologically correct' characteristics. Previous related data abstraction schemes are discussed briefly for comparison purposes. In the case of interrelated tasks and corresponding bodies of 2D maps, the latter can facilitate the recognition of associations and the inference of rules, important components of Machine Learning.

Keywords: dimension-reduction, data abstraction, machine learning, 2D maps, cluster analysis.

1. 2D Maps in the Context of Machine Learning

Machine Learning is that aspect of machine behavior which enables a machine to improve its own performance in some task domain through acquisition of knowledge from experience in that and related domains. That view of Machine Learning is illustrated in Figure 1, patterned after Figure 1.1 of [1]. Some critical and ill-understood matters happen at the step labeled "Acquiring Experience" and this discussion is concerned primarily with that step.

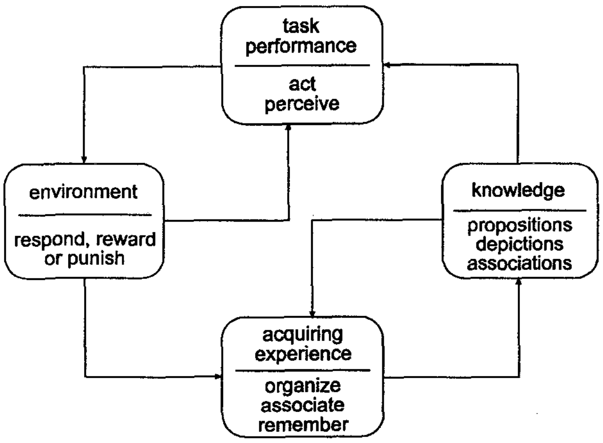

Figure 1: Elements of machine learning. (Adapted from [1].)

In particular, this discussion is concerned with the task of making sense out of large bodies of high dimensional data.

Two issues are of primary importance, one being the manner in which a body of data points are distributed or grouped within a single data space and the other the issue of establishing associations between data groupings in different domains. Conventionally, the first activity is usually described in terms of self-organization, or unsupervised learning, or cluster analysis, whereas the second is of fragmented nature and is referred to variously as rule inference, or categorization, or assessing the weight of evidence and so on. The first activity, the so-called cluster analysis, is a difficult one, more difficult than is generally realized. And the second activity becomes difficult if the first is not done well.

It is the intent of this present brief discussion to argue that it is possible and useful to map multidimensional data into 2D representations in robust and meaningful manner. In that reduced-dimension form, the shapes and locations of the data groupings can then be managed accurately and groupings can then be linked to data groupings in other related data spaces. In other words, it is argued that 2D maps can facilitate the important task of ' Extended Cluster Analysis' in Machine Learning. However the quality of the dimension-reduction mapping is of the utmost importance.

A new dimension-reduction procedure is described in Section 2. It is well specified computationally, and has attractive characteristics, the most important of which is that the topology of the original space seems to be preserved in some manner. Clusters in high-dimensional space seem to remain singly connected groupings in 2D and data points which were close in high-dimensional space remain close in 2D space. 'Topological correctness' seem to be maintained to some degree in both coarse grained and fine grained manner. Some other previously investigated dimension-reduction methods are discussed briefly in Section 3. These include accounts of two other avenues of investigation devised by us. In addition, the role of 2D maps in enabling Machine Learning is discussed schematically and briefly in Section 4 for illustration purposes.

2. The Distance-Ratio Conserving Mapping

The objective is to be able to abstract high-dimensional data sets and display the essence of that information in 2D. The task may also be understood in terms of 'Nonlinear Principal Component Analysis'. Many methods have been proposed by us and by other researchers. In this section we describe one specific preferred method and present some illustrative results. The mapping is carried out with a conventional multilayer feedforward net, trained in conventional Back-Propagation manner. The criterion for the training of the mapping is that the ratio of a point-to-point distance in full dimension to the same distance in reduced dimension be the same for all pairs of points. In practice this is difficult to achieve and in practice what is done is to minimize the variance of all such ratios. The equations for learning of the parameters in the dimension-reduction net are shown in Figure 2.

$$\|\mathbf{o}_p - \mathbf{o}_{p'}\| = \sqrt{\sum_{k=1}^{K}(o_{kp} - o_{kp'})^2}, \qquad \|\mathbf{x}_p - \mathbf{x}_{p'}\| = \sqrt{\sum_{i=1}^{I}(x_{ip} - x_{ip'})^2}$$

$$E = \frac{2}{P(P-1)}\sum_{p=2}^{P}\sum_{p'=1}^{p-1}\frac{\|\mathbf{o}_p - \mathbf{o}_{p'}\|^2}{\|\mathbf{x}_p - \mathbf{x}_{p'}\|^2} - \left[\frac{2}{P(P-1)}\sum_{p=2}^{P}\sum_{p'=1}^{p-1}\frac{\|\mathbf{o}_p - \mathbf{o}_{p'}\|}{\|\mathbf{x}_p - \mathbf{x}_{p'}\|}\right]^2$$

$$\Delta w_{kj} = -\eta\frac{\partial E}{\partial w_{kj}} = -\eta\frac{4}{P(P-1)}\sum_{p=2}^{P}\sum_{p'=1}^{p-1}\left(\frac{\|\mathbf{o}_p - \mathbf{o}_{p'}\|}{\|\mathbf{x}_p - \mathbf{x}_{p'}\|^2}\frac{\partial}{\partial w_{kj}}\|\mathbf{o}_p - \mathbf{o}_{p'}\|\right)$$
$$- \frac{8}{P^2(P-1)^2}\left(\sum_{p=2}^{P}\sum_{p'=1}^{p-1}\frac{\|\mathbf{o}_p - \mathbf{o}_{p'}\|}{\|\mathbf{x}_p - \mathbf{x}_{p'}\|}\right)\sum_{p=2}^{P}\sum_{p'=1}^{p-1}\left(\frac{1}{\|\mathbf{x}_p - \mathbf{x}_{p'}\|}\frac{\partial}{\partial w_{kj}}\|\mathbf{o}_p - \mathbf{o}_{p'}\|\right)$$

$$\frac{\partial}{\partial w_{kj}}\|\mathbf{o}_p - \mathbf{o}_{p'}\| = \frac{(o_{kp} - o_{kp'})[o_{kp}(1 - o_{kp})o_{jp} - o_{kp'}(1 - o_{kp'})o_{jp'}]}{\|\mathbf{o}_p - \mathbf{o}_{p'}\|}$$

$$\Delta w_{ji} = -\eta\frac{\partial E}{\partial w_{ji}} = -\eta\frac{4}{P(P-1)}\sum_{p=2}^{p}\sum_{p'=1}^{p-1}\left(\frac{\|\mathbf{o}_p - \mathbf{o}_{p'}\|}{\|\mathbf{x}_p - \mathbf{x}_{p'}\|^2}\frac{\partial}{\partial w_{ji}}\|\mathbf{o}_p - \mathbf{o}_{p'}\|\right)$$
$$- \frac{8}{P^2(P-1)^2}\left(\sum_{p=2}^{P}\sum_{p'=1}^{p-1}\frac{\|\mathbf{o}_p - \mathbf{o}_{p'}\|}{\|\mathbf{x}_p - \mathbf{x}_{p'}\|}\right)\sum_{p=2}^{P}\sum_{p'=1}^{p-1}\left(\frac{1}{\|\mathbf{x}_p - \mathbf{x}_{p'}\|}\frac{\partial}{\partial w_{ji}}\|\mathbf{o}_p - \mathbf{o}_{p'}\|\right)$$

$$\frac{\partial}{\partial w_{ji}}\|\mathbf{o}_p - \mathbf{o}_{p'}\| = \frac{1}{\|\mathbf{o}_p - \mathbf{o}_{p'}\|}\times$$
$$\left\{\sum_{k=1}^{K}(o_{kp} - o_{kp'})[o_{kp}(1 - o_{kp})w_{kj}o_{jp}(1 - o_{jp})x_{ip} - o_{kp'}(1 - o_{kp'})w_{kj}o_{jp'}(1 - o_{jp'})x_{ip'}]\right\}$$

Figure 2: The equations for learning of parameters in dimension reduction net.

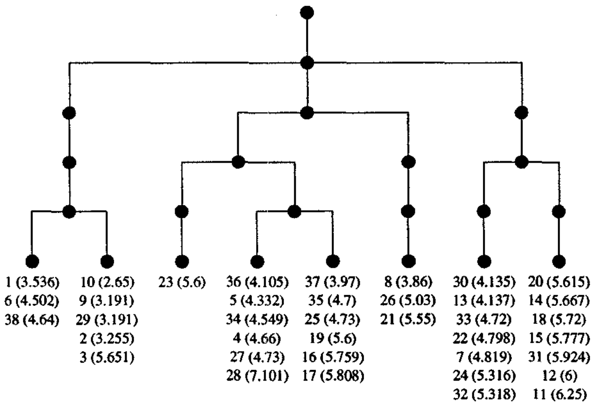

Figure 3: A Hierarchical Cluster Structure Obtained for A Body of Semiconductor Data.

In this brief discussion we content ourselves with illustrating the dimension-reduction procedure with a body of semiconductor data. Namely a set of semiconductor compounds are each described in terms of five feature values, these being the molecular weight of chemical formula, dimensions of unit cell, size of largest cation and value of band-gap. These values are scaled to lie approximately between +1 and -1. A hierarchical K-means clustering procedure results in a tree structure of clusters such as that shown in Figure 3. The full-dimension data are subsequently subjected to distance-ratio conserving mapping, implemented using a neural network trained in accordance with the equations exhibited in Figure 2, and the plot of Figure 4 shows that data points which belonged to any one cluster in original full-dimensioned data space also tend to lie close to each other in the reduced-dimension 2D space. Detailed examination of the data point id's confirm this understanding of the results. In addition, there is a sense of the coarse grained inter-cluster spacings. The 2D map shows that the designation of the distribution in terms of clusters is somewhat arbitrary, the distribution actually being quite loose. Nevertheless the relative spacing of the clusters and of the patterns are made more evident in this manner. That type of knowledge can be very useful in understanding how multivariate data items might be managed.

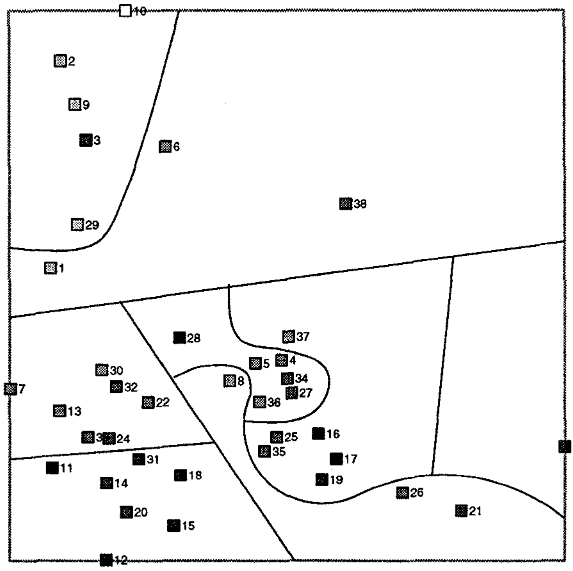

Figure 4: Clusters and reduced-dimension plot. For a body of semiconductor compounds data originally described in 5D.

Another reduced-dimension plot is shown in Figure 5. The data displayed are for another problem entirely but the plot can be used to illustrate the use of the 2D plots for providing a quick grasp of the sensitivity of the position of the data points to changes in the feature values in the original high-dimensional space.

3. Other Dimension -Reduction Procedures

The conventional supervised learning paradigm of neural-networks using multilayer feedforward network architectures and gradient descent learning can be viewed as performing a mapping from a set of vectors in N dimensional space to a corresponding set of points in K dimensional space, with K=2 for obtaining 2D maps. From that viewpoint, it would seem there is nothing new or implausible in seeking to carry out dimension-reduction, and to some extent this is true. However in conventional supervised learning we know both the inputs and the outputs for a set of data points (the training set) and the criterion for learning the mapping parameters is that the estimated outputs be as close as possible to the expected values (the targets). The objective function is the mean of the square of the difference between the two sets of values, the known expected values of the outputs and the estimated values given by the mapping. The learned mapping minimizes that objective function and should be valid for all new samples as well as the points of the training set. What is new in dimension-reduction is that there are no target values and the methods differ depending on the objective function selected.

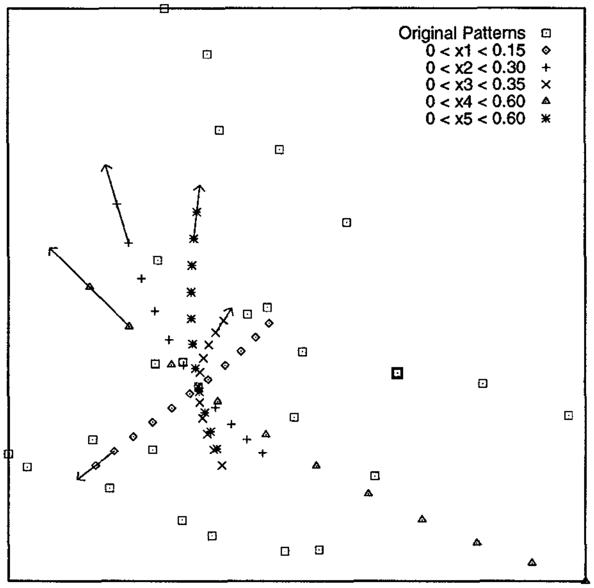

Figure 5: Display of Sensitivity of Data Point Position to Changes in Value of Original Features.

In the following we list some dimension-reduction procedures and describe them in a uniform neural networks framework .

a. The Karhunen-Loeve Transformation [2] [3]: Consists of learning a full-dimensioned N to N linear mapping; the linear mapping coefficients are learned iteratively by minimizing the mean of the square of the off-diagonal components of the data covariance matrix, the result being that the covariance matrix is diagonalised. Finally only those few components of the new representation for which the diagonal components are large are retained. (Note in passing: Of course this is not the traditional approach for carrying out the K-L transform, and is inefficient computationally, but it is a neural network procedure capable of finding the eigenvectors and eigenvalues of the covariance matrix.)

b. The Auto-Associative Mapping [4] [5]: Actually this is a conventional supervised learning system.

Given a set of N input values one attempts to map that into a set of identical N output values with use of a multilayer feedforward neural net. The challenge come from the requirement that a single set of network parameters suffice for all sets of input values, that is for all pattern vectors in a body of data. In addition there is a constraint that one of the internal layers only have K nodes. The outputs of those nodes are taken to provide the reduced-dimension representation. For reduction to 2D, K=2.

c. The Variance Constraint Mapping [6]: A neural network is trained to map a set of points in N dimensional data space into a corresponding set of points in K dimensional space subject to the constraint that the trace of the data covariance matrix be conserved.

d. The Distance-Ratio or Metric Ratio Conservation Mapping [7]: A general nonlinear mapping is learned on the condition that for any and all pairs of points, the ratio of the original distance to the transformed distance be the same for all the pairs. This being impossible in general, the constraint used is that the variance of those values be as small as possible. [This is the method of preference at this juncture of our explorations of the various methods].

e. The Equalized-Orthogonal Projection with Relaxation Mapping [8] : A general nonlinear mapping based on the condition that the data covariance matrix in the transformed space be as close as possible to the identity matrix multiplied by a constant.

f. The SAMANN Mapping [9] [10] : A general nonlinear mapping patterned after the Sammon algorithm, based on the constraint that all inter pattern distances be conserved in the reduced-dimension space. This is a more demanding condition than that required of the Metric-Ratio Conserving Mapping.

g. The Self-Organizing-Map [SOM] or feature map approach [11] [12] : In this method a set of pattern vectors are mapped onto a 2D grid of nodes on the basis of similarity to reference vectors belonging to the grid points. The reference vectors are selected randomly at first from the same domain as the data vectors. For each data vector, after an association between that data vector and a node reference vector is made, the receiving node and neighboring nodes are made to be more like the incoming vector. In this way the entire grid becomes self-organised in a clustering manner. In other words the patterns rely on the grid nodes to locate and associate with other data vectors similar to it. In this way the data vectors gather themselves into 'clusters' organised spatially in the space of the grid nodes. These clusters transform the reference vectors and in so doing partition the grid space in a graduated manner. The grid is usually a 2D grid and that is how this relates to dimension reduction. The relative placement of the clusters may be governed by the topology of the original data but may also be influenced greatly by the initial random assignment of the reference vectors, by the volume of data available and by the size and density of the nodes available in the feature map.

h. GTM-The Generative Topographic Mapping: The mapping task is cast in the form of determining the data distribution in reduced-dimension space which is consistent with the data distribution in the original data space. In the model and narrative developed in [13], one starts with a grid of points in the latent-dimension space and assumes that the set of points x(i) can be mapped in a one-to-one manner into a set of corresponding points y(i) in original data space, the mapping being in the form of a linear weighted sum of nonlinear basis functions defined in latent space. The values of the weights are learned adaptively. The points y(i) do not cover all of data space. To remedy this situation one erects Gaussian distributions about all the y(i). In this manner one obtains an estimate of the distributions at any and all points t in data space. The variance of the Gaussian may be learned adaptively. Now one addresses the fact that the distribution values at the actual data points t(n) are known, and proposes that the Bayes' relationship can be used to estimate what the corresponding distribution should be in latent space. The reestimated distribution should be the same as the original estimated value. This condition is used for learning of the linear weights and of the spread of the Gaussians centered about the points y(i). The underlying theory seems to depend on three sets of interacting assumptions all worthy of further study.

4. Use of 2D Maps in Machine Learning

In Machine Learning research literature there are many articles proposing how knowledge might be represented, managed and used. For example there are suggestions that it would be useful to differentiate between deep knowledge and shallow knowledge, or that it is important to have some knowledge in the form of causal rules and others in the form of heuristic rules, and yet other forms of knowledge in the form of frames. In other directions of differentiation, one distinguishes between various forms of rule inference. All of these concerns may or may not be important but the theme of this discussion is that perhaps the main issue is none of those topics. We suggest that one of the most important issues is how to handle the multidimensional aspects of data items and of rules.

It is interesting to note that several diverse research disciplines tend to feel that they can handle quite well the task of describing the design and implementation of an Intelligent Machine System capable of learning, and they are almost correct. But they all tend to falter in the matter of handling the complexity brought about by the multidimensional nature of the knowledge items involved. This difficulty surfaces in different ways. In Fuzzy Set technology the difficulty is manifest in the issue of whether it is ever possible to have a universally valid General Extension Theorem for combining single variable membership function values; in inference and search, over a body of rules, the different antecedents of the various rules are dealt with sequentially rather than in parallel, resulting in great computational complexity; Dempster-Shafer theory deals with the same issue under the guise of procedures for 'combining evidence'; in probabilistic causal reasoning that same issue manifests itself in the fact that very large numbers of different inference paths need to be scanned, with each path associated with products of conditional probabilities and 'priors' of dubious accuracy.

In pattern recognition and neural networks, the approach is quite different. The many dimensions and the many feature values are treated in parallel. This does not mean that the basic mathematical complexities have been banished by magic but rather that the issue is now cast in a form such that the combinatorial complexities can be readily curtailed by appeal to experience. In other words, in a multidimensional feature space, each point in that space might be thought of as corresponding to a specific combination of feature values. In the pattern recognition or neural network approach, the data points are allowed to self-organize and one finds that, in nearly all cases, not all eventualities happen with equal frequency. In fact, data points tend to appear in groups, and even these can be described in abstracted form, 2D maps being one of these abstracted forms.

In conclusion, we claim that 2D maps can represent effectively all the knowledge entities known to be of importance in intelligent systems technology including models of processes, heuristic rules, analytically expressed causal relationships and so on. In addition, 2D maps are well suited to dealing with the effects of experience, in accepting new data and in updating knowledge structures. 2D maps also can support the processes of chaining and scanning, but in efficient parallel manner rather than in sequential manner. Because of all this, the matter of 2D map formation is worthy of careful study.

References

[1] Langley, P., 1996. *Elements of Machine Learning*, Morgan Kaufman Publishers, Inc., San Francisco, CA.

[2] Fukunaga, K. and W. L. G. Koontz, 1970. Application of the Karhunen-Loeve expansion to feature selection and ordering, *IEEE Transactions on Computers*, Vol.19, pp.311-318.

[3] Fukunaga, K., 1990. *Introduction to Statistical Pattern Recognition, Second Ed.*, Academic Press, NY.

[4] Kramer, M., 1991. Nonlinear principal component analysis using auto-associative neural networks, *AIChe*, vol. 37, pp. 233-243.

[5] Pao, Y.H., 1996. Dimension reduction, feature extraction and interpretation of data with network computing. *International Journal of Pattern Recognition and Artificial Intelligence*, World Scientific Publishing Co., Vol.10, pp.521-535.

[6] Pao, Y.H. and C.Y. Shen, 1997. Visualization of pattern data through learning of nonlinear variance- constrained dimension-reduction mapping, *Pattern Recognition*, Vol.30 (in press).

[7] Pao, Y.H., Z. Meng, S.R. LeClair and B. Igelnik, 1997. Validation of distance ratio constrained dimension-reduction displays of multidimensional data, submitted to *Engineering Applications of Artificial Intelligence*, also technical report, AI WARE Inc., Beachwood, OH.

[8] Meng, Z. and Y.H. Pao, 1997. Visualization and self-organization of multidimensional data through equalized orthogonal projection with relaxation, *Internal Technical Report*, Electrical Engineering and Applied Physics, Case Western Reserve University, Cleveland, OH.

[9] Sammon, J.W., Jr., 1989. A nonlinear mapping for data structure analysis, *IEEE Transactions on Computers*, vol. C-18, pp.401-409.

[10] Mao, J. and A.K. Jain, 1995. Artificial neural networks for feature extraction and multivariate data projection, *IEEE Transactions on Neural Networks*, Vol. 6, pp.296-316.

[11] Kohonen, T., 1982. Self-organized formation of topologically correct feature maps, *Biological Cybernetics*, Vol.43, pp.59-69.

[12] Kohonen, T., 1995. *Self-organization Maps*, Springer, NY.

[13] Bishop, C.M., M. Svenen and C.K.I. Williams, 1997. G.T.M: the generative topographic mapping, accepted for publication in *Neural Computation*.

[14] Meng, Z., Y.H. Pao, and C.Y. Shen, 1996. Optimization with dimension reduction through learning of nonlinear variance constrained mapping, *Proceedings of the Adaptive Distributed Parallel Computing Symposium*, Dayton, OH, pp.208-217.