ELSEVIER

Journal of Alloys and Compounds 279 (1998) 22–29

Journal of ALLOYS AND COMPOUNDS

Materials discovery via topologically-correct display of reduced-dimension data

Y.H. Pao[a,*], Z. Meng[a], S.R. LeClair[b], B. Igelnik[a]

[a]*Case Western Reserve University and AI WARE Inc., Cleveland, OH 44106, USA*
[b]*Materials and Manufacturing Directorate, Air Force Research Laboratory, WPAFB, OH, USA*

Abstract

Authors describe and demonstrate a 'ratio-conserving' mapping procedure for attaining reduced-dimension representations of multidimensional data. This procedure has a theoretical basis for topological correctness in that the ratio of the metrics in the two representations is maintained constant throughout. It is also demonstrated that comparing the reduced-dimension depiction of data with the results of clustering and spanning tree operations in the full-dimension space can validate such reduced-dimension mappings. Demonstrations are carried out using a body of semiconductor data, with five independent variables and one dependent variable.

Keywords: Dimension reduction; Materials design; Materials property prediction; Neural networks

1. Introduction

This paper is concerned with the task of reducing the complexity of bodies of data. It is suggested that a special kind of dimension-reducing mapping might be useful for such complexity reduction purposes. The mapping is topologically correct in detail, in the sense that the mapping is able to ensure a minimum of variance in the mapping ratio over the entire pattern space in question. This mapping procedure is described and demonstrated with a body of semiconductor data.

It is difficult to make sense out of a large body of multidimensional data. If we imagine each record to be a point in the multidimensional data space, it is difficult to visualize how these points are distributed in that space, whether the data points are in tight clumps or uniformly distributed, or some of this and some of that in different regions of that space.

There are two approaches to the reduction of the complexity of the data, one, by clustering so that in the case of tightly grouped clusters, a large number of patterns (data points) can be represented quite adequately by the cluster center or prototype. This methodology is widely practice. The K-Means and ISODATA algorithms of pattern recognition are widely known and widely practice [1]. Echoes of that approach can be found in neural-net computing clothed in terms such as unsupervised learning [2] and associative memories [3]. The intent of such practices may be considerably different, but at some stage comparable methodologies are employed and comparable beneficial results are obtained.

In another approach, reducing the dimensionality of the data reduces complexity. This might be accomplished by exploring and discovering that for some of the features, the feature values do not vary significantly from one pattern to another. In that case, the feature values can be represented by a constant average value for all the patterns in the data set. In the Karhunen–Loeve approach [4], a linear coordinate transformation is carried out so that the pattern representation is in a space spanned by the eigenvectors of the covariance matrix.

Those coordinates, which have low variance, are represented by constant values and the active part of the representation is reduced. The so-called auto-associative neural network mapping [5,6] may be thought of as a nonlinear extension of the Karhunen–Loeve philosophy.

It is interesting to think of Kohonen's Feature Map approach [7] as constituting a combination of the two methods for reducing complexity. The method attempts to map multidimensional data points onto a space of lower dimension and attempts to group together points which were close to each other in the original data space.

It might be noted however that even though Kohonen has emphasized a desire to achieve topologically correct

*Corresponding author.

0925-8388/98/$19.00
PII: S0925-8388(98)00608-2

mapping [8], it would seem that what is achieved in the Feature Mapping procedure is a looser measure, namely all that is rather similar in the original space should be also rather close in the reduced-dimension space. Nevertheless, as is known, the Feature Map methodology is interesting and useful.

2. Topological correctness based on ratio conservation

In this discussion we present a methodology for achieving a mapping based on the criteria that all inter-pattern distance ratios be conserved (as nearly as possible) in the mapping. It would seem that the degree of topological correctness is extended to a finer detail than in the Feature Map approach. The idea of ratio conservation dictates that if the ratio of the distance between patterns p and p' in the full dimension space to the distance between the same two patterns in reduced-dimension space is *r*, then all the ratio values for all other possible pattern pairs be that same value. In the event that it is not possible to have all such ratios exactly the same, then the variance of all the ratio values over that set of pattern pairs should be as small as possible.

We propose that a ratio-conserving mapping can be learned with use of a multilayer neural network. The equations for learning the network parameters for achieving such a mapping are exhibited in Fig. 1.

This approach to dimension reduction was applied to a body of semiconductor data listed in Table 1, provided by Dr. A.G. Jackson of US Air Force Air Force Research Laboratory, Wright–Patterson Air Force Base, OH 45433-6523. In that table each semiconductor compound is described in terms of five supposedly independent variables, namely the electronic band gap for excitation to the conducting states, two crystal unit cell parameters, the atomic weight of the compound, and the 'radius' of the anion. For the present purposes neither the accuracy of the values nor the appropriateness of the variables are of concern. This is a typical format in which a body of multidimensional data might be presented.

Also, for demonstration purposes, a value of the density of the compound is associated with each entry of five values of independent variables. Again, in this discussion, there is no guarantee of the accuracy of the values as provided. This body of data is being used as provided for the purposes of demonstrating this new ratio-conserving mapping, and for interpreting the results of that mapping.

One result of such a mapping is shown in Fig. 2 where the five-dimensional data are exhibited in two dimensions. In Fig. 2, each data point is represented by a square icon

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

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

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

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

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

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

Fig. 1. Equations for learning the network parameters for achieving a ratio-conserving mapping with use of a multilayer neural network.

Table 1
Characteristics and properties of some semiconductor compounds (used as a representative body of data for illustration purposes, courtesy of Dr. A.G. Jackson, Wright Laboratory, WPAFB, OH)

n	Compound	Gap	*a*	*c*	*w*	*r*	*ρ*
1	ZnS	3.9	3.823	6.261	97.434	53	3.536
2	AlN	6.2	3.11	4.98	40.99	25	3.255
3	ZnO	3.3	3.251	5.209	81.369	22	5.651
4	$AgGaS_2$	2.638	5.751	10.238	241.718	53	4.66
5	$CuGaS_2$	2.43	5.351	10.47	197.388	53	4.332
6	$LiIO_3$	4	5.481	5.171	181.836	22	4.502
7	Se	1.7	4.361	4.954	78.96	66	4.819
8	GaS	2.5	3.586	15.496	101.784	53	3.86
9	SiC	6	4.359	4.359	40.09	29	3.191
10	$Si0_2$	8.4	4.9134	5.4052	60.078	22	2.65
11	Te	0.33	4.457	5.939	236.55	82	6.25
12	AgI	2.8	6.473	6.473	234.77	126	6
13	CuCl	3.17	5.405	5.405	98.993	77	4.137
14	CuI	2.95	6.042	6.042	190.44	96	5.667
15	InSb	0.23	6.479	6.479	236.55	89	5.777
16	$AgGaSe_2$	1.8	5.981	10.865	335.51	66	5.759
17	$AgInSe_2$	1.2	6.099	11.691	286.798	66	5.808
18	InAs	0.36	6.268	6.479	189.79	71	5.72
19	$CdGeAs_2$	0.57	5.943	11.217	334.97	71	5.6
20	GaSb	0.72	6.095	6.095	191.47	89	5.615
21	InSe	1.25	4.002	24.946	193.76	66	5.55
22	InP	1.35	5.868	5.868	145.77	59	4.798
23	Ag_3AsS_3	2	10.8	8.69	494.792	53	5.6
24	GaAs	1.4	5.653	5.653	144.71	71	5.316
25	$CuGaSe_2$	1.7	5.606	11.006	242.468	66	4.73
26	GaSe	2.021	3.747	23.91	148.68	66	5.03
27	$CuInS_2$	1.53	5.489	11.101	242.468	53	4.73
28	HgS	2.1	4.145	9.496	232.654	53	7.101
29	β-SiC	2.26	4.359	4.359	40.09	29	3.191
30	GaP	2.3	5.45	5.45	100.69	59	4.135
31	ZnTe	2.3	6.101	6.101	192.97	82	5.924
32	ZnSe	2.7	5.667	5.668	144.33	66	5.318
33	CuBr	2.91	5.69	5.69	143.449	82	4.72
34	$CdGeP_2$	2.91	5.74	10.776	246.93	59	4.549
35	$ZnSiAs_2$	1.74	5.606	10.88	243.43	71	4.7
36	$ZnGeP_2$	2.05	5.463	10.731	199.9	59	4.105
37	$CdGa_2S_4$	3.05	5.568	10.04	380.096	53	3.97
38	$LiNbO_3$	4	5.148	13.86	147.842	22	4.64

and labelled with the chemical formula of the compound. The gray level color of the icon is used to provide a qualitative depiction of the density of the compound.

A critical issue accompanying such dimension reduction mappings is the matter of uniqueness and faithfulness of such resulting representation. Or more accurately, how are these reduced-dimension representations to be interpreted, understood or utilized. This issue is of interest because reduced-dimension displays do provide appealing visualizations of data but may be misleading without a rigorous theoretical basis for such mappings. As mentioned previously, the Feature Map is primarily concerned with making sure that those patterns which are close in full dimensional space remain close in reduced-dimension space, addressing topological correctness in an overall average manner.

In the next section of this discussion, we display the results of hierarchical clustering of the patterns in full dimensional space and show how the results of the clustering operation helps in the validation and interpretation of the reduced-dimension mapping.

3. Validation of the ratio-conserving mapping

It is possible to try to organize the original data into clusters with a procedure such as the K-means algorithm. In that procedure one specifies the number of clusters desired, starts off with K arbitrarily chosen points in pattern space as the cluster centers and processes all data points by assigning each and every one to the nearest cluster center, with distance measured according to the Euclidean Distance metric. The position of a cluster center is updated by taking it to be the average of all points in that cluster. The algorithm ends when all data points have been assigned to cluster centers and there is no further change in cluster centers or cluster assignments. In that

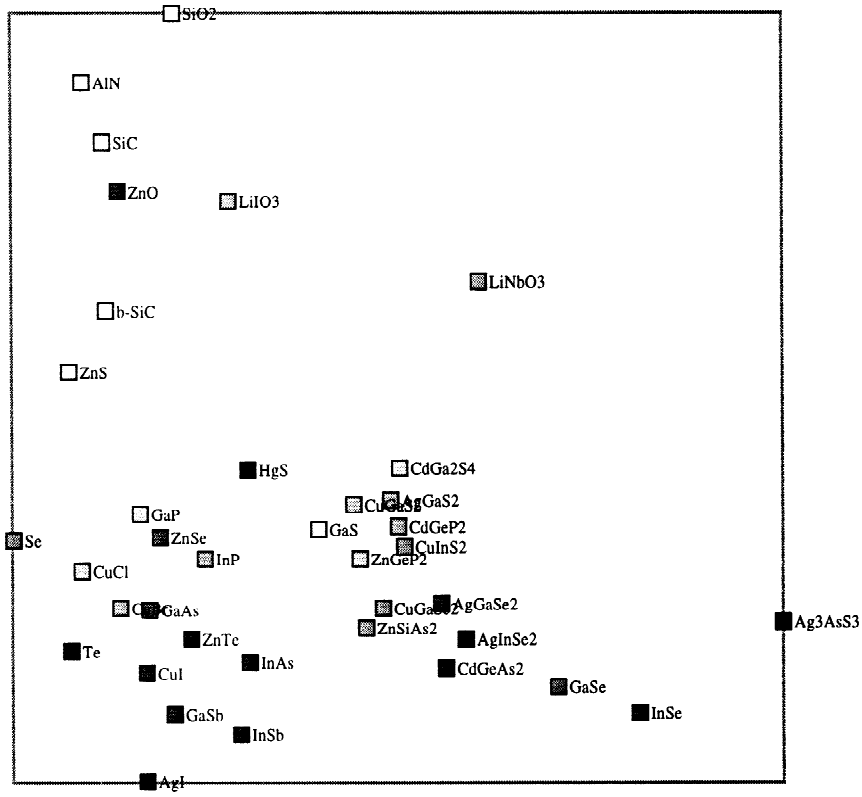


Fig. 2. The various five-featured semiconductor patterns mapped onto a 2D display. (Nearly topologically correct mapping).

procedure there is no mention of cluster radius. If there are too many data points in any one cluster, that cluster might be subdivided into several other clusters, using the same K-means algorithm but only for the patterns in that one cluster in question. One obtains, in that case, a hierarchical cluster tree which might be used for rapid processing of a very large number of patterns.

The cluster structure is in the nature of a tree. The result for K-means hierarchical clustering for the semiconductor data of Table 1 is illustrated in Fig. 3 involving structures of three, four, six and eight clusters.

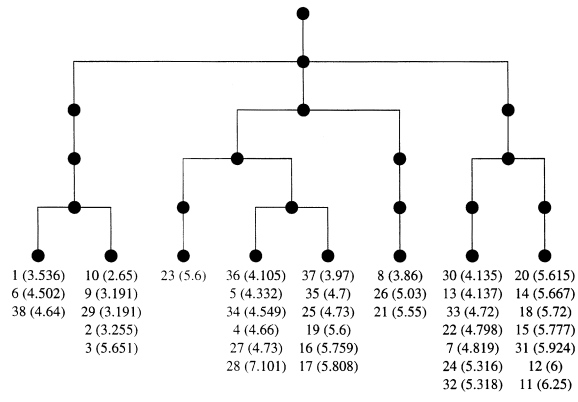


Fig. 3. Hierarchical cluster structure of semiconductor data.

The combined effects of clustering and reduced-dimension mapping are shown in Fig. 4 for the above structures. In each of the cases, the inter-cluster boundaries are drawn in accordance with the cluster membership assignments of Fig. 3. In most of the cases, the inter-cluster boundaries could be taken to be straight lines, there being a lack of data to justify greater detail in the placing of the boundaries. In reality, there is no basis for expecting those borders to be straight lines.

It is very satisfying that compounds which were found to be ‘close’ to each other in SD space are indeed grouped together in 2D space, or more specifically all the patterns in each and every cluster fall within singly connected areas in reduced-dimension space. In other words, the clusters validate the reduced-dimension mapping, and the latter provides a visualization of inter-pattern and inter-cluster distances.

A further sense of validation may be attained by linking all the patterns together with a minimal spanning tree. Such a tree is displayed in Fig. 5(a). It was obtained with use of the Greedy Algorithm [9]. The linkages constitute a tree and not a graph. The algorithm consists of starting off with a link representing the shortest of all pair-wise inter-pattern distances, this instantiates one link and two nodes

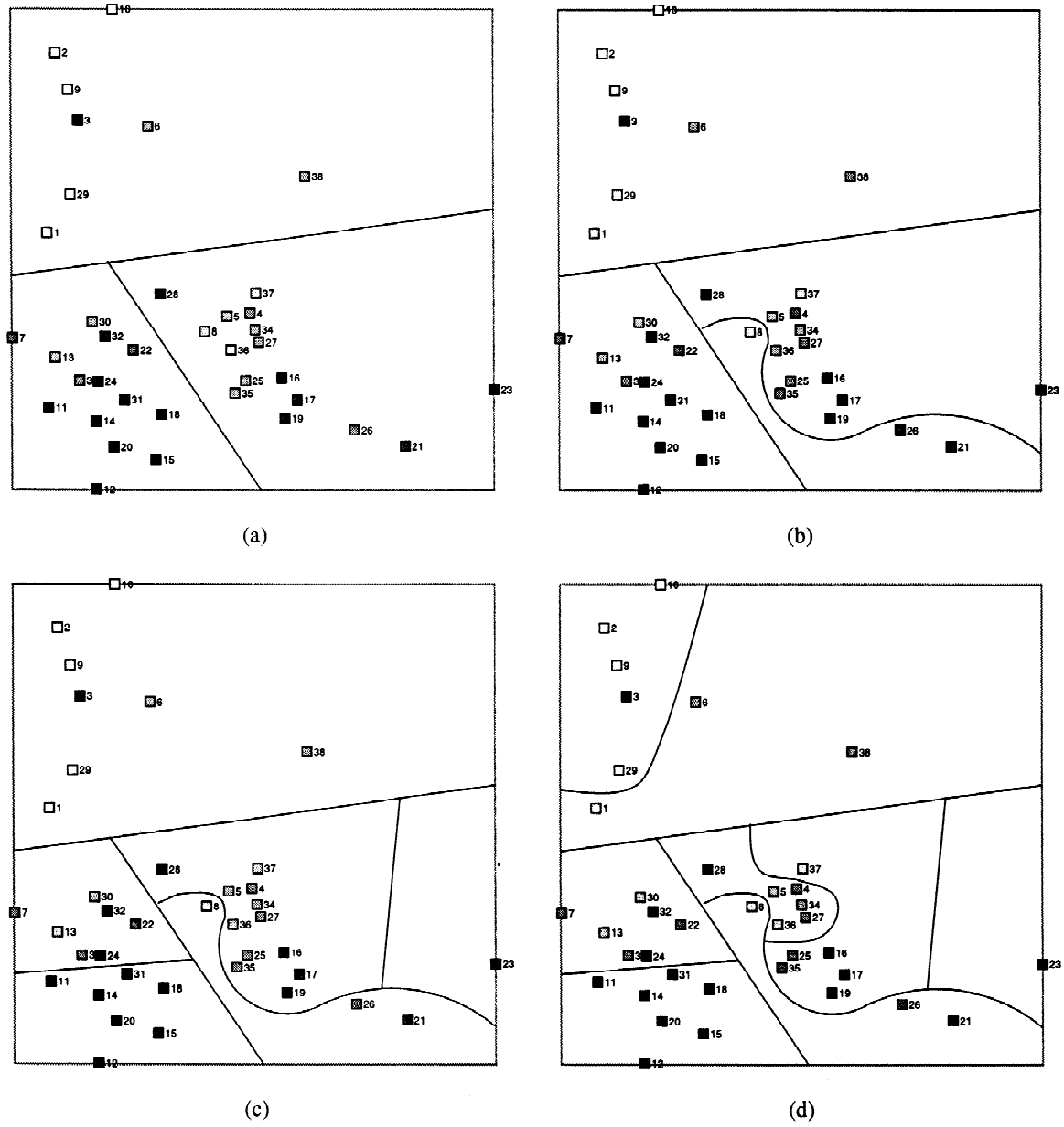


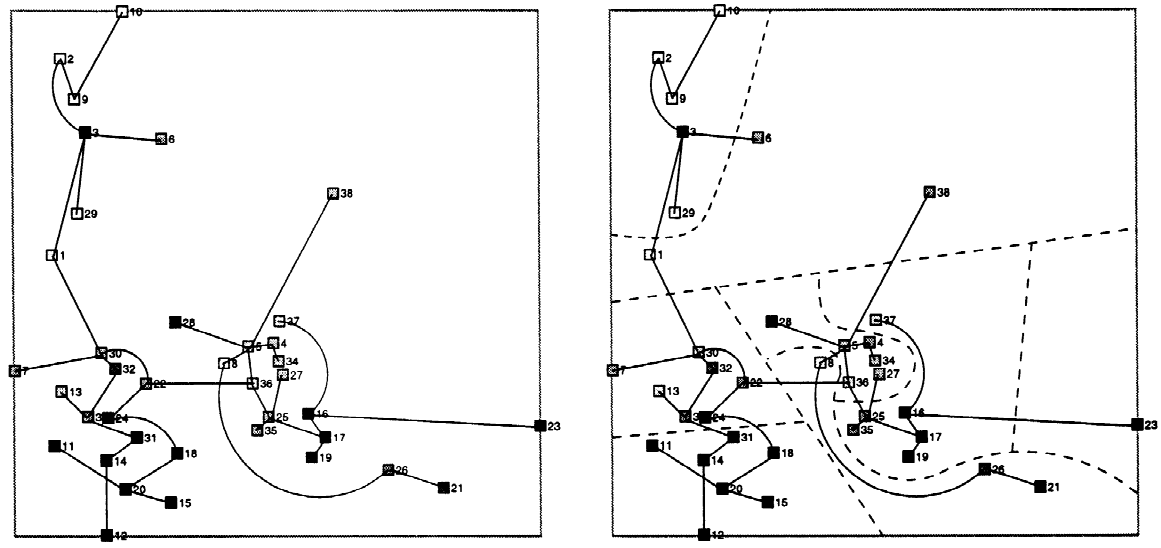
Fig. 4. Cluster and reduced-dimension plots. For visualization of inter-pattern and intra-cluster relationships.

or patterns. The next step is to find the next shortest link which can be connected to one of the nodes which have already been instantiated. The algorithm continues in this manner until all patterns have been activated and drawn as nodes in the tree, taking care to avoid the formation of closed loops. The complexity of this procedure is of the order of N^2 and is not necessarily the most efficient [9], but it suffices for the present purposes.

All three modes of visualization are combined in Fig. 5(b). The inter-cluster linkages of the spanning tree provides additional reassurance of the fidelity of the ratio-conserving mapping. Patterns which are close in SD are indeed close in 2D and those inter-cluster links can serve as the basis for error recovery in search, or for guidance in exploration and discovery.

4. Visualizing multidimensional data

The display of Fig. 2 is shown again in Fig. 6 with the square icon identified in terms of the density of the compound instead of the chemical formula. The cluster boundaries are also plotted. It is seen that the clusters do



(a) (b)

Fig. 5. A minimal spanning tree perspective. (a) The minimal spanning tree for the 38 locations in SD space; (b) A more complete visualization of the data including cluster boundaries.

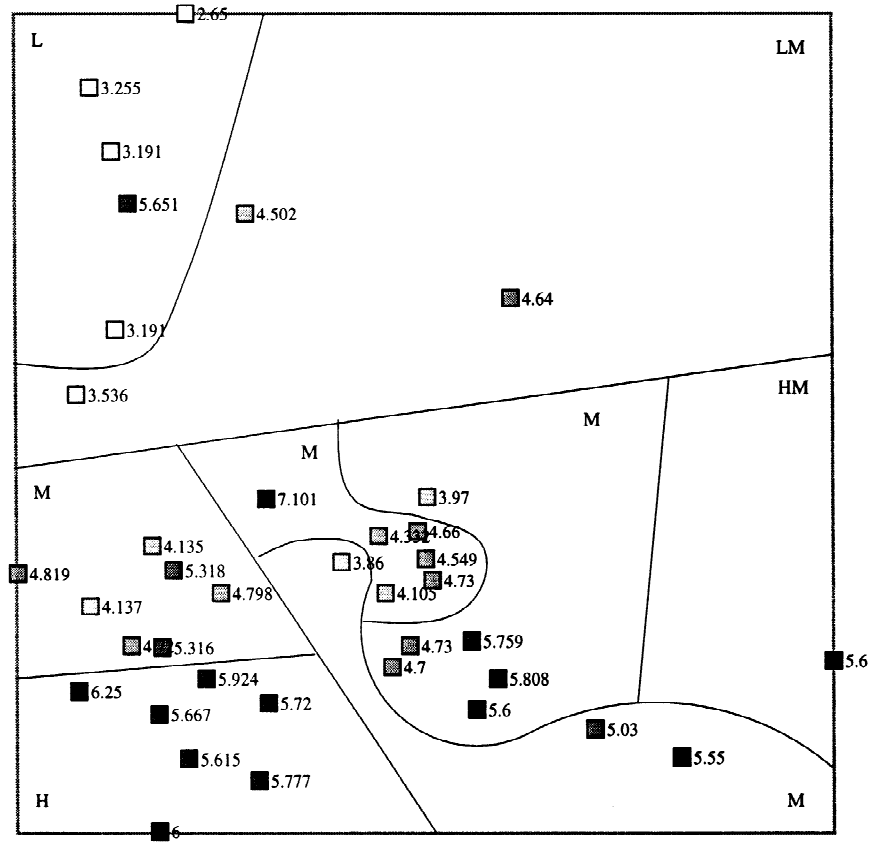


Fig. 6. Use of reduced-dimension plot for obtaining an overview of variation of density. (L, LM, M, HM and H denotes the regions of Low, Low-Moderate, Moderate, High-Moderate and High values of density values respectively).

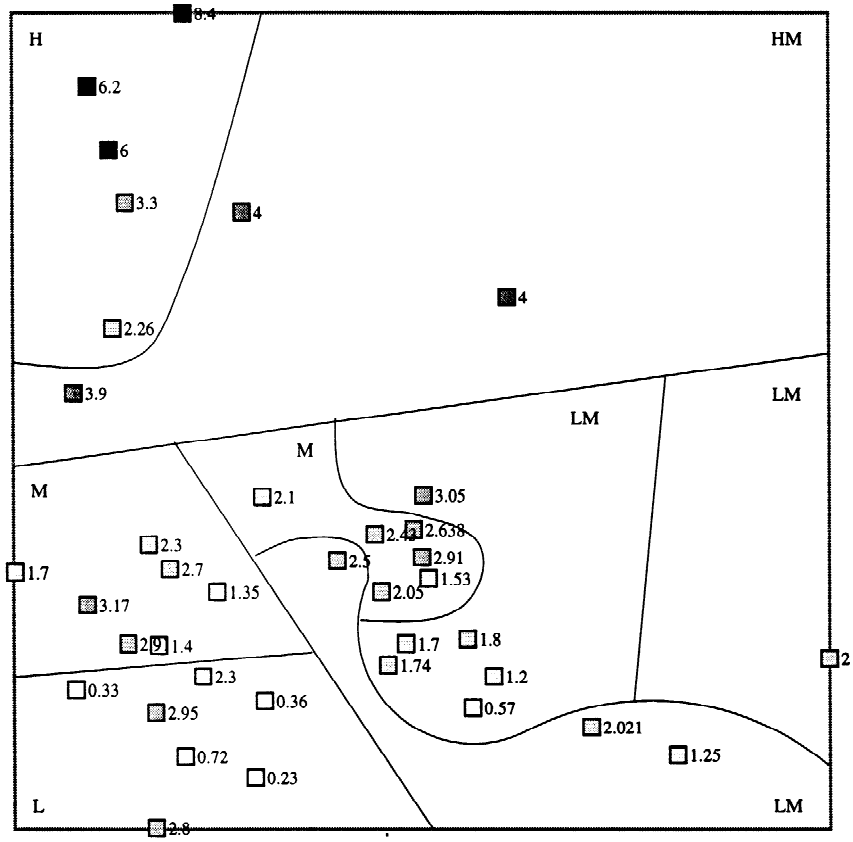


Fig. 7. A fuzzy correlation between cluster membership and band gap values as viewed in reduced dimension.

not correspond to crisp categories as far as density values are concerned. Instead it might be said that the clusters correspond to several fuzzy sets with designations of High, High-Moderate, Moderate, Low-Moderate and Low values of density. Given the reduced-dimension representation, membership functions can be learned for each and all of the clusters, or functional approximations of the density values may be learned for most of the clusters, the ones which have some representative body of data. Given any new compound, it would be possible to estimate what the density value is likely to be.

A similar plot is shown in Fig. 7 for band-gap values. However, the band-gap is supposedly an independent variable and not a dependent variable such as density. The fact that some regions have low band-gaps and others higher values suggest that there is significant functional dependence between band-gap and the other four supposedly independent variables.

5. Summarizing remarks

A dimension-reducing mapping is described in this paper. This mapping is quite unusual in that it is learned with the use of a neural net and a focus on maintaining the change in the metric as uniform as possible throughout the pattern space. This condition provides a robust theoretical basis for this topologically correct mapping. A clustering operation in the original full-dimensional space yields a partitioning of that original full-dimension space. It is gratifying to see that all the clusters map into singly connected areas in the 2D space. This validates the reduced-dimension mapping. It would seem that patterns which are close in SD are indeed also close in 2D. But it also seems that not only distances are maintained proportionately but a sense of orientation and many-body configurations are also retained. The reduced-dimension plot is useful in that it provides a sense of how the clusters relate to each other. That visualization makes many other data-mining operations readily feasible.

References

[1] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.
[2] Y.-H. Pao, Adaptive Pattern Recognition and Neural Networks, Addison–Wesley, Reading, MA, 1989.

[3] G.A. Carpenter, S. Grossberg, Appl. Opt. 26 (1987) 4919–4930.
[4] K. Fukunaga, W.L.G. Koontz, IEEE Trans. Computers 19 (1970) 311–318.
[5] M. Kramer, AIChE J. 37 (1991) 233–243.
[6] Y.-H. Pao, International Journal of Pattern Recognition and Artificial Intelligence 10(5) (1996) 521–535.
[7] T. Kohonen, Self-organization Maps, Springer, New York, 1995.
[8] T. Kohonen, Biol. Cybern. 43 (1982) 59–69.
[9] D. Harel, Algorithmics: The Spirit of Computing, Addison–Wesley, Reading, MA, 1987.