Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis A.Ultsch, H.P. Siemon Institute of Informatics University of Dortmund Po. Box 500 500 D-4600 Dortmund 50, W.Germany

Abstract

In this paper we describe experiments with self organizing feature maps that have been implemented on a transputer system. The use of feature maps for clustering is investigated and it is shown that a naive application of Kohonen's algorithm, although preserving the topology of the input data, is not able to capture clusters. A new method, called U-matrix, is proposed which is capable to classify correctly all examples. First experiments with medical data of high dimensionality show a high correlation with expert clustering of the data.

1. Introduction

Since its introduction in 1982 organizing feature maps have been used for a variety of applications including robot programming, modelling of VLSI production, electrophoresis diagnostic medical picture processing and others [Kohonen 82, Bertsch/Dengler 87, Keller/Fogelman 88, Ritter/Schulten 86]. All these applications rely on the ability of Kohonen's algorithm to adapt itself suitably to the structure of a high dimensional data space. In this paper we investigate the use of Kohonen's maps for exploratory data analysis, in particular for cluster analysis.

In chapter two we describe the construction of self organizing feature maps with Kohonen's method. We show that an direct usage of Kohonen maps is not suited best for the purpose of clustering. In chapter three a novel method how Kohonen maps can be used in order to allow exploratory data analysis is proposed. Chapter four gives first results for a nontrivial example. 2. Feature Maps for Clustering

The following is a short description of Kohonen's algorithm to construct feature maps[Kohonen] 84]. Let S be a subset of \mathbf{R} , I = {x₁, ..., x_d, x_i \in Sⁿ} the input data and d(x,y) a vector norm defined on Sⁿ, U a lattice of n dimensional vectors (called units) with a mapping n: U x U \rightarrow R⁺ called neighborhood. The mapping $f:S^n \rightarrow U$: $f(x) = u_{ij}$, iff $d(x,u_{ij}) = \min \{d(x,u), u \in U\}$ is called feature map iff the lattice U is constructed using the following Training Algorithm. U is also called a Kohonen net with dimension o. Training Algorithm:

For t = 0 to T do (learning epoch): for each $x \in I$ do(learning step): for each $u \in U$:

 $u(t) = \begin{cases} u(t-1) & \text{if } n(f(x),u) > b(t) \\ u(t-1) + \eta(t) * (x - u(t-1)) & \text{if } n(f(x),u) \le b(t) \end{cases}$ where b(t) is a decreasing real valued function

with $b(0) \le 0$, b(T) = 0 and $0 \le \eta(t) < 1$, a function on t which decreases with time t. In the following we will use as lattice U a rectangular two-dimensional array of vectors, i.e

 $U=\{u_{ij}\in \mathbb{R}^n \mid 1 \le i, j \le o \in \mathbb{N}\}\$ and as neighborhood function: $n(u_{ij}, v_{kl}) = \sqrt{(j-k)^2 + (j-l)^2}$. The algorithm has been implemented on a transputer system with 17 processors [Ultsch/Siemon 89]. Special implementation techniques yielded an overall learning speed of 2.4 MCUPS (mega connection updates per second). A special graphical interface allowed us to observe the training algorithm. See [Siemon/Ultsch 90] for details.

One of the properties of a Kohonen net should be that the topological (neighborhood) relationships among the input are reflected as faithfully as possible in the arrangement of the corresponding units in the lattice [Ritter/Schulten 86, Kohonen 84]. Cluster analysis groups "similar" objects i.e. objects that have common features into disjunct subsets called clusters [Deichsel/Trampisch 85]. If the property, that a Kohonen map is topology conserving holds, then clusters in the input data, i.e. subsets of data, that are close neighbours in Rⁿ should be mapped onto the lattice U with the same close relationship. Any clusters in Rⁿ should also appear in the lattice U of lower dimensionality. In order to investigate the clustering capabilities of a Kohonen map, we used data generated in the following way: We choose 4 points (ABCD) in the three-dimensional unit cube, that constructed a tetraeder with edge-length of .96. (see figure 1a) For each of these four points ten random

vectors were generated, having a length in the range from .1 to .4 with a mean length of .2. The random vectors were added to the edge points of the tetraeder in order to get four separated data clusters. All vectors were scaled such that they fitted into the three-dimensional unit-cube (i.e. S^n $= I^{3}$). Figure 1b shows the the data clusters.



Figure 1: a) the tetraeder,

b) the data clusters A clustering algorithm should separate the data set into exactly four subsets and indicate, that the four subsets are dissimilar. To test this, a Kohonen net of dimension 64 was used and trained with the 40 data points, leaving out the vectors A to D. Figure 2a shows the initial distribution of f(x)of each data point; Figure 2b depicts the Kohonen net after 50.000 learning steps. This type of diagram is called coordinate matrix in the sequel.





Figure 2 Tetraeder coordinate matrix: a) initial, As it can bee seen, the net arranges the clusters into the different corners of the lattice U. The

distances among the data points, however, are evenly distributed. If their membership in the original clusters would not be known, no clustering could be detected. Other experiments with hexaeders and octaeders showed the same behaviour [Ultsch/Siemon 89].

3 The U-Matrix Method

We have seen in the last paragraph, that coordinate matrices are not directly suited to detect clusters of data in the input space. In this paragraph we will describe a method called U-matrix method, that allows to get a more suitable picture of the vector distribution.

For a unit uii in the lattice U four different types of distances to its immediate neighbours can be defined as follows: $dx(i,j) = d(u_{ij},u_{i+1j})$, $dy(i,j) = d(u_{ij},u_{i+1j})$, $dxy(i,j) = d(u_{ij},u_{i+1},u_{i+1})$ and $dyx(i,j) = d(u_{ij+1},u_{i+1},u_{i+1})$. We propose the following method to combine the four distances combined into one matrix $\in \mathbb{R}^{2n-1\times 2n-1}$ as follows:

ned mile one n	incentive at	40 10110			
U-matrix		2j-1	2j	2j+1	
2i-1		dz(i-1,j-1)	dy(i-1,j)	dz(i-1,j)	
2i		dx(i,j-1)	du(i,j)	dx(i,j)	
2i1		dz(i, j-1)	dy(i,j)	dz(i,j)	

306

Where dz(i,j) = 0.5*(dxy(i,j)+dxy(i,j) and du(i,j) may be arbitraryly chosen. This can be used, for example, to indicate the correspondence of this entry to unit u_{ij} . This diagram we call *unified* distance matrix or short U-matrix. For each unit u_{ij} in U there is a corresponding element du(ij)in the U-matrix. The entries next to du(i,j) contain the distances dy, dy and the diagonal distances at their geometrical correct places. The diagonal distanced dxy and dyx are represented on the diagonal elements of the U-matrix as the arithmetic mean of dxy and dyx. The U-matrix contains therefore a geometrical correct approximation of the vector distribution in the Kohonen net. To get a visual impression on how this distribution is, we propose to display the U-matrix in three dimensions i.e. display its elements as a height over a grid that corresponds to the lattice. This display has "valleys" where the vectors in the lattice U are close to each other and "hills" or "walls" where the vectors in the lattice U have a larger distance.

Figure 3a shows the U-matrix for the tetraeder data. As it can be seen, the map is now divided into four different regions, corresponding to the four data clusters. The regions are separated by a "wall" i.e. a large vector to vector distance, indicating a dissimilarity in the input data. The U-matrix for hexaeder data shows the same effect (see figure 3b). In this case the data is separated into six different regions.





Figure 3: a) U-matrix of tetraeder data

b) U-matrix of hexaeder data

To explore the properties of a n-dimensional data space we propose to interpret the landscape of the U-matrix. We call this method the *U-matrix method* [Ultsch/Siemon 89]. Given a n-dimensional data set I the question is, whether there exists any structure in the form of subsets of data that are very similar to each other:

- 1) Construct a Kohonen map on a two dimensional array U;
- 2) Calculate the U-matrix as described above;
- 3) Map the input data onto the U-matrix by using the feature map f;
- 4) Interprete the U-matrix as follows:

If a subset C of input data I falls into a valley in the U-matrix, then C is a cluster in I, i.e. C contains similar vectors. If the input data-vectors are neighbours in the U-matrix, they are close to each other in \mathbb{R}^n . If different depressions are separated by walls or are geometrically far apart, then there is a large dissimilarity among the different clusters. The relative location of clusters in the U-matrix reflect their (dis-) similarities in \mathbb{R}^n . The higher the walls between clusters, the more dissimilar are the clusters in \mathbb{R}^n .

4. Application of the U-matrix method

To test our method with a data set that has, first, more than three dimensions and, second, stems from real life data we have used the data set in a standard text book on cluster analysis [Deichsel/Trempisch 85]. The data consists of blood measurements. Each of the 20 data vectors contains 11 different blood values. According to medical diagnosis, the data can be groped into eight healthy patients, three patients with metabolical acidosis, one patient with a cerebral deficiency five patients with respriratory acidosis and three patients with lactacidosis.

The different components of the data vectors are measured in different scales. In order to apply a vector metric for all components the range of the components has to be the same, otherwise components with a larger absolute range would dominate the vector metric. To solve this problem

4

we have used the so called z-transformation on each vector's components: subtract the coordinate mean from the coordinate value and divide by the standard deviation of this coordinate (see e.g. [Deichsel/Trampisch 85]). Figure 5a shows the coordinate matrix and figure 5b the U-matrix for this data.





Figure 5: a)Coordinate matrix of blood data 5. Discussion

b) U-matrix of blood data

The U-matrix method allows a Kohonen map to group n-dimensional data into clusters of similar data. This grouping of the data corresponds with the medical diagnosis of the patients. More than a ready made clustering algorithm we see our method as a tool for the inspection of high

dimensional data. The algorithm can be thought of as a mapping from \mathbb{R}^n to a nonlinearily flattened two-dimensional surface such that interesting topological relations are conserved. The regions where the two-dimensional surface is bent and the amount of bending is represented topologically correct in the U-matrix. Other methods to depict the properties of a high dimensional data space like Chernoff's faces of Kleiner Hartigan trees rely on the human ability to compare and abstract pictures of faces respectively trees [Barnett 81]. Our method uses geometrical closeness as a measure for similarity. For the separation of different groups (clusters) of data our method uses the third dimension in the form of walls. Fist experience with the method suggest a high correlation with expert diagnosis of data. In a current test series we will experiment with high dimensional data about drinking water quality and with blood data of patients with a mangelkrankheit.

6. **Čonclusion**

In this paper we have investigated the use of Kohonen's self organizing feature maps for exploratory data analysis. A naive application of Kohonen's algorithm, although preserving the topology of the input data is not able to show clusters inherent in the input data. A new method, called U-matrix, is proposed. This method is capable to classify correctly all artificially generated data. Moreover experiments with data of high dimensionality stemming from the area of medicine show a high correspondence with expert diagnosis of the data. As a first result this encourages the application of Kohonen's algorithm for the use in exploratory data analysis.

References

[Barnett 81] Barnett, V., (ed.): Interpreting Multivariate Data, John Wiley and Sons, 1981.

[Bertsch/Dengler 87]Bertsch,H.,Dengler,J.: Klassifizierung und Segmentierung medizinischer Bilder mit Hilfe der selbstlernenden topologischen Karte,in:Paulus,E.(ed.):DAGM-Symposium Mustererkennung, Springer Informatik Fachberichte 149,1978, 16 -170.

[Deichsel/Trampisch 85]Deichsel, G., Trampisch, H.J.: Clusteranalyse und Diskriminanzanalyse, Gustav Fischer, Stuttgart.

[Keller/Fogelman 88] Keller, M., Fogelman-Soulié, F.: Topological maps and their applications to electrophoresis image recognition, Proc. Neuro-Nimes, November Nimes, France, 1988, pp 403 -413.

[Kohonen 82] Kohonen, T.: Self Organized Formation of Topologicaly Correct Feature Maps, Biol. Cybern. 43, 59-69. [Kohonen 84] Kohonen, T.: Self-Organisation and Associative Memory, Springer Verlag, 1984.

[Ritter/Schulten 86] Ritter, H., Schulten, K.: Toplology conserving mappings for learning motor tasks, Proc. Conf. Neural Networks for Computing, Utah, 99 -106.

[Siemon/Ultsch 90] Siemon, H.P., Ultsch, A.: Kohonen Networks on Transputers: Implementation and Animation.

[Ultsch/Siemon 89] Ultsch, A., Siemon, H.P.: Exploratory Data Analysis: Using Kohonen Networks on a Transputer, Research Report in Computer Science, University of Dortmund, December 1989.

308