In Proceedings of SSGRR 2000, Int. Conf. on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet, L'Aquila, July 31 - August 6. Scuola Superiore G. Reiss Romoli, 2000. (Proceedings on CD-ROM)

Methods for exploratory cluster analysis

Samuel Kaski, Janne Nikkilä, and Teuvo Kohonen

Abstract— **When beginning the analysis of a new data set of which very little is known a priori, the first step is to explore the data. This paper presents new methods for this preliminary data mining phase: for detecting, visualizing, and interpreting cluster or density structures of the data. The groundwork for the visualizations is a nonlinear ordered map display constructed using the Self-Organizing Map algorithm. The detected structures can be interpreted using linear local factors extracted from the nonlinear map. In a case study using a collection of patent abstracts the methods have, for instance, detected a cluster of neural networks patents not previously distinguished by the international patent classification system.**

Keywords— **Clustering, Exploratory Data Analysis, Multivariate Analysis, Self-Organizing Map (SOM)**

I. Introduction

In this paper we address the question of how to begin the analysis of a large high-dimensional data set when very little is known of the data a priori. For simplicity, the data set is assumed to consist of real-valued vectorial items having no time dependence.

The first step in any analysis should aim at gaining as much understanding of the intrinsic dimensions of the data distribution as possible, and if the data is high-dimensional this task is far from trivial. We introduce methods for visualizing and interpreting cluster structures of the data, to help in its interactive exploration.

II. Map of the data space: Groundwork for visualization

Properties of high-dimensional data sets have been visualized using a wide variety of methods ranging from glyphs and scatterplots to multidimensional scaling and hierarchical trees. Our alternative is a graphical map display, a regular grid on which closeby locations represent similar data items. The map is defined by associating domains in the input space with points of the display in an orderly fashion. This is done using the Self-Organizing Map algorithm. Intuitively the graphical map display then corresponds to a nonlinear but smoothed network in the input space. Each data item is projected onto the location of the map display for which the network is closest to the item.

The map displays have several desirable properties. (1) The feature map metaphor, interpretation of the display as a map of the essential features of the data set, makes the display intuitive. (2) Even complicated high-dimensional data distributions can be visualized satisfactorily with a nonlinear map that is fitted to the data distribution. (3) The map display provides an ordered groundwork for visualizing the clusteredness of the data, for example, and the values of the original variables in different subdomains. (4) Whereas the interpretation of local properties of the original high-dimensional data set may be difficult, the map may be interpreted more easily as will be demonstrated in Sec. V-C. The interpretations provide insights of the data as long as the map fits the data well, which may be tested easily if necessary. Items (3) and (4) are the main topics of this paper.

The Self-Organizing Map algorithm [1], [2] is a method for constructing such mappings. Since there already exists extensive literature on the SOM [3], we will present here the algorithm and some of its properties only briefly. The graphical display consists of a regular grid of map units. There is a model vector $\mathbf{m}_i$ associated with each unit i. The model vectors have the same dimensionality as the data items[1], and they define the points of the input space that are associated with the locations of the display. As a result of the SOM algorithm described below, the model vectors approximate local data domains in an orderly fashion: model vectors of closeby map units are similar. The "elastic network" corresponding to the map becomes fitted to the data distribution. The network goes "through the data", as close to the data points as possible, while remaining organized ("smooth"). The point density of the nodes approximates the density of data.

The values of the model vectors are computed in an iterative process which may be either stochastic or batch-type. Here we will consider a batch-type solution for a given data set $\{\mathbf{x}_k\}$. In the so-called Batch Map algorithm [2] all data vectors $\mathbf{x}_k$ are first projected to the map by finding their closest model vectors, indexed by c_k:

$$c_k = \arg\min_i \|\mathbf{x}_k - \mathbf{m}_i\| \; . \tag{1}$$

The distance measure is usually the normal Euclidean distance, but other kinds of measures can be used as well. The new value of each model vector, used in the next iteration, is computed in two steps. First, the centroid of the data projected onto each map unit is formed. This is a usual vector quantization step:

$$\bar{\mathbf{x}}_j = \frac{\sum_{c_k=j} \mathbf{x}_k}{n_j} \; , \tag{2}$$

where n_j is the number of the data samples projected to the jth map unit. Second, the new value $\mathbf{m}_i^*$ of the model vector of the ith unit is computed as a weighted average of the centroids of neighboring units. This is similar to

The authors are with the Neural Networks Research Centre, Helsinki University of Technology, P.O. Box 5400, FIN-02015 HUT, Finland. E-mail: Samuel.Kaski@hut.fi, Janne.Nikkila@hut.fi, Teuvo.Kohonen@hut.fi

[1]Here we will consider only vectors although more complicated models may be used as well.

smoothing:

$$\mathbf{m}_i^* = \frac{\sum_j n_j h_{ji} \bar{\mathbf{x}}_j}{\sum_j n_j h_{ji}} , \quad (3)$$

where h_{ji} is the smoothing kernel or *neighborhood function* that is a monotonically decreasing function of the distance between the ith and jth map unit on the map grid. The smoothness of the discrete manifold formed of the model vectors can be controlled by the width of the neighborhood function relative to the size of the map.

When interpreting the SOMs in Sec. V we will utilize concepts from traditional linear principal component analysis. We will therefore next consider its relation to SOM more closely.

Principal components describe the *global* distribution of the data in terms of linear combinations of the original variables or, equivalently, projections to straight lines in the data space. The first principal component is defined as the projection of the data to the line for which as much of the variance of the data is preserved as possible. The second principal component is the projection to the line that is orthogonal to the first one and preserves the remaining variance maximally, and so on.

Principal curves are generalizations of the principal components, defined to be smooth curves that go through the "center of the data". Each point of the curve is the average of all data points projecting to it. This definition is not applicable as such for discrete data sets but some kind of smoothing has to be applied, resulting in various algorithms depending on the type of smoothing.

The principal curves [4] and surfaces [5] are closely related to the SOM, and they provide additional insight on the relation between the SOM and principal components. A discretized principal surface, together with a certain kind of smoothing, would be essentially a SOM [6], [7]. (It should be noted, however, that the SOM was derived independently eight years earlier than the principal curves.) The local direction of the "elastic network" formed of the SOM may be interpreted as an estimate of the direction of *local principal variation* in the data. The so-called *local factors* computed from the SOM will be utilized to characterize the direction in data exploration in Sec. V-C.

Note: The dimensionality of the graphical map display is usually selected low, preferably two, to enable versatile visualization. It is clear, of course, that all properties of a high-dimensional data set cannot be visualized using such a low-dimensional display, and sometimes higher-dimensional (e.g. 3D) grids are used to preserve the topology of the data space better. If the intrinsic dimensionality of the data set is high, the SOM is still capable of representing the data by "folding" itself in the data space, but then the direction of the SOM is no longer as meaningful. Fortunately, however, natural data sets are often *locally* relatively low-dimensional (they have a low fractal dimensionality), and even when they are not, the higher-dimensional domains can be easily detected by estimating the local dimensionality of the data at each map location, and visualizing it on top of the map display. The estimate of the local dimensionality functions as an indicator of the confidence in the results.

III. Case study: maps of text document collections

A database of text collections will be used to demonstrate the methods. The collection contains 5982 patent abstracts from the following groups of the international patent classification system: optical computing devices, electrical digital data processing, analogue computers, and hybrid arrangements.

The document collection has been processed using a new variant of the SOM developed for textual databases, called WEBSOM [8], [9]. Statistical indicators of the word occurrences (histograms of words) were used to represent documents, which could then be treated as real vectors.

The goal of the WEBSOM is to organize a free-form text document collection onto a document map on which documents having similar content are located near each other. A document map display provides an overview of the document collection, and it can be used as a browsing and exploration interface. Additionally, the WEBSOM can be used as a content-addressable search engine in information retrieval applications, whereby the results of the searches can be visualized on the document map display. The latest version of WEBSOM has been streamlined to make computation of very large maps feasible; the largest map so far contains about 7,000,000 patent abstracts. In this paper we will use a smaller map with a simpler preprocessing stage to demonstrate the novel tools for interpreting and visualizing the cluster structure of the data.

Before computing the SOM the patent abstracts were preprocessed by removing from the total vocabulary of 9856 words the ones that occur less than 50 times, and words belonging to a stop list of 1335 words. The remaining vocabulary consisted of 988 words. The documents were encoded as word histograms, in which each bin corresponds to the frequency of occurrence of a certain word in the document. Each bin was additionally weighted by an indirect measure of its importance, an entropy-based measure of the unevenness of the distribution of the word in the four classes of the patent classification system. The weighted histograms were used as inputs to a SOM of the size of 595 units.

The neighborhood function of the SOM had a Gaussian shape and its standard deviation in the end of the computing process was 2 distance units (one unit equals to the distance between two neighboring map units on the hexagonal map grid). The resulting map is analyzed in detail in the following sections.

IV. Visualizing cluster structures in the data

A. Classical methods

There exist two categories of classical statistical tools that are especially useful in exploring high-dimensional data sets: clustering and projection algorithms, respectively. Classical clustering algorithms produce a grouping of the data according to a chosen criterion. For instance, within-cluster variance may be maximized while minimizing between-cluster variance. Projection methods, on the other hand, represent the data points in a lower-dimensional space that can be visualized. Different projection methods minimize different cost functions; the classical principal component analysis, for instance, aims at preserving a maximal proportion of the variance in the original data.

Both kinds of algorithms are valuable in data exploration. Clustering compresses the large amount of data into representative categories, while projection methods make it possible to explore high-dimensional proximity relations between data points. However, both have well-known disadvantages: the (nonlinear) projection methods are formidably slow for large data sets and therefore an additional reduction in the number of data points is needed. The most prominent disadvantages of most clustering algorithms is that the fine structure within the clusters is lost, and it is difficult to judge whether the extracted clusters are "natural" properties of the data or merely one possible partitioning of the data set.

B. SOM-based methods

The Self-Organizing Map can be used for both nonlinear projection and visualization of cluster information. Even cluster borders can be sought by studying the relations between neighboring models in the map. More commonly, however, the SOM is used for the *visualization of the cluster (or density) structures of the data.* A measure of the local clusteredness of the data is computed for each map location and plotted as gray levels on the map. The so-called U-matrix method [10] visualizes the distances of the model vectors attached to neighboring map grid locations. Since it is known that the density of the model vectors reflects the density of the data points, the distances between neighboring model vectors reflect the density.

Alternatively, the map can be colored so that the perceptual similarity between the colors reflects the metric relations of the high-dimensional data items as faithfully as possible [11]. Different clusters then attain different colors and the changing color emphasizes changes in between the clusters.

In data exploration, these methods have the advantage compared to classical clustering algorithms that the fine structures in the probability density of the data becomes visible. If the final goal is to extract clusters, the SOM displays can be used for exploring the clusteredness visually before extracting the clusters automatically.

C. A new, improved method for detecting and visualizing gaps between clusters

Data exploration has usually two different goals: obtaining an overview, a general understanding of the nature of the data, and detecting novel, unexpected and potentially interesting properties of the data. The U-matrix is a proper tool for the former goal although without additional processing the visualizations may be noisy (cf. Fig. 1**a**). In this section we will consider the latter goal.

In cluster analysis it is often implicitly assumed that the data consist of a set of clearly discernible, isolated and compact clusters. Approaches that try to extract such clusters may, however, give unintuitive results since real-valued measurements are usually noisy and there rarely exist clear distinct clusters. Moreover, high-dimensional data are usually sparse.

If there exist, however, isolated clusters or "gaps" in the data distribution, they then reflect very interesting properties of the data. Paradoxically, the key to clustering seems to be to regard the clusters as unexpected properties in the data!

We have developed a method that can be used for sensitive detection of gaps between higher-density domains. Such gaps are candidates for cluster edges, or otherwise interesting domains of data. The method searches for domains in which (1) the density is low, and (2) the gradient of the sample density changes its direction. For each model vector $\mathbf{m}_i$ of the SOM, we first compute the vector

$$\mathbf{g}_i = \mathbf{c}_i - \mathbf{m}_i \tag{4}$$

pointing from the model vector $\mathbf{m}_i$ to the centroid $\mathbf{c}_i$ of the closest points around $\mathbf{m}_i$, defined by

$$\mathbf{c}_i = \frac{1}{N_i} \sum_{\mathbf{x}_k \in I_i} \mathbf{x}_k \; . \tag{5}$$

Here I_i denotes the set of indices of the N_i data points that are closest to $\mathbf{m}_i$. The number of data points N_i is chosen to be a certain fraction, say 80%, of the data projected to the neighborhood of the map unit i. The search for the closest data points may be speeded up by considering only data projected within the neighborhood of unit i on the map, instead of the whole data set, to get an approximation.

The vector $\mathbf{g}_i$ points to the direction in which the density of the data increases, or towards the closest cluster if the density around $\mathbf{m}_i$ is very low. Within a domain of constant density the magnitude of $\mathbf{g}_i$ is zero, and the magnitude increases as the gradient of the density increases. On the other hand, the lower the density the larger the magnitude of $\mathbf{g}_i$. The magnitude of the *difference* between $\mathbf{g}_i$ and $\mathbf{g}_j$ of neighboring map grid locations is therefore a working detector of between-cluster domains. Intuitively speaking, the difference for two neighboring units is large only if (1) there is a sharp minimum of density in between the model vectors $\mathbf{m}_i$ and $\mathbf{m}_j$, or (2) the model vectors are close to different clusters.

In summary, the measure we will use for detecting between-cluster domains is

$$G_{ij} = \|\mathbf{g}_i - \mathbf{g}_j\| \; . \tag{6}$$

The value of G_{ij}, converted into gray shades, has been plotted in between each pair of neighboring map grid points i and j in Figure 1b. Dark stripes on the display depict detected cluster borders.

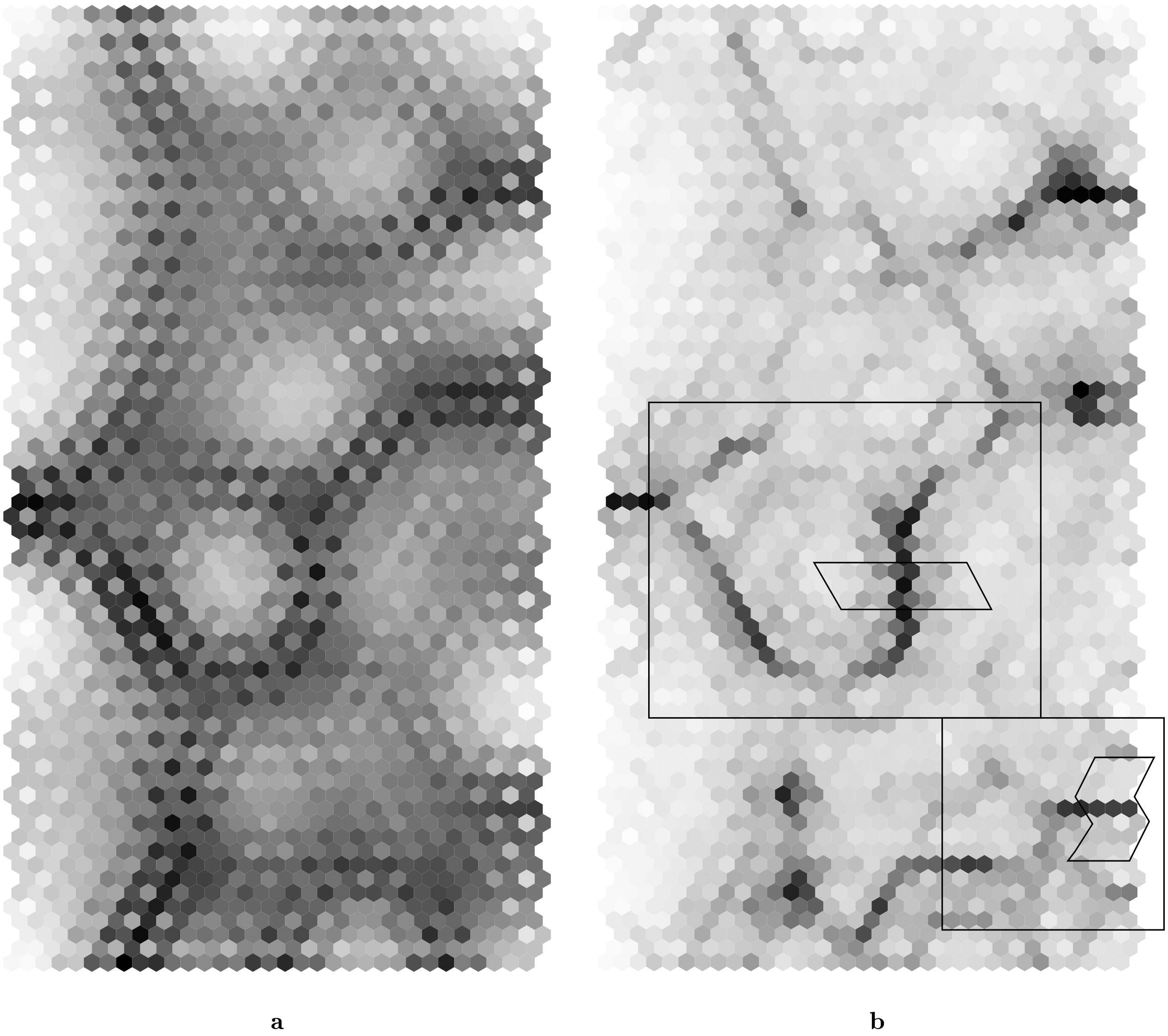

Fig. 1. Methods for visualizing cluster structures on the SOM groundwork. Both displays consist of one hexagon for each grid point on the 17 by 35 unit SOM, and one hexagon in between each pair of neighboring grid points on a regular hexagonal lattice. The data set consisted of 5982 patent abstracts. **a** U-matrix. Light shades denote denser distributed model vectors and dark shades big jumps in model vector values, respectively. **b** The new method for detecting gaps between clusters, denoted by darker shades. The areas outlined with the black lines will be analyzed in more detail in Sec. IV-D.

Connection to the U-matrix. Note that G_{ij} in (6) can be written as

$$G_{ij} = \|(\mathbf{m}_j - \mathbf{m}_i) - (\mathbf{c}_j - \mathbf{c}_i)\| = \|U_{ij} - (\mathbf{c}_j - \mathbf{c}_i)\| \ , \quad (7)$$

where U_{ij} is the value of the U-matrix at the corresponding location. The U_{ij} measures the distance along the elastic network fitted to the data by the SOM algorithm. The elastic network is a smoothed representation of the data; it is known [2] that the equilibrium values of the model vectors are in the centroids of the data projected to them or their neighbors. The second term $\mathbf{c}_j - \mathbf{c}_i$ is computed from non-smoothed local data. If it cancels the smoothed value then the data is fairly regular and the value of G_{ij} is small. If the smoothed difference is not canceled out, however, the residual reflects irregularity or unexpectedness in the data. Such irregularity is typically caused by changes in the density around the edges of clusters.

D. Case study: Clusters in patent data

In this preliminary case study we studied the patent abstracts in the clustered area that was most saliently distinguished from its surroundings on the map shown in Figure 1. This cluster and its surroundings have been magnified in Figure 2. Randomly selected titles of patents mapped to nodes on a sample area around the edge have been highlighted. It turned out that almost all of the patents within the area treated neural networks, whereas almost none of the patents in the neighboring areas did.

Neural networks, being a relatively new research area, did not have a class of their own in the international patent classification system at the time the abstracts had been classified. Starting from year 2000 there is a new

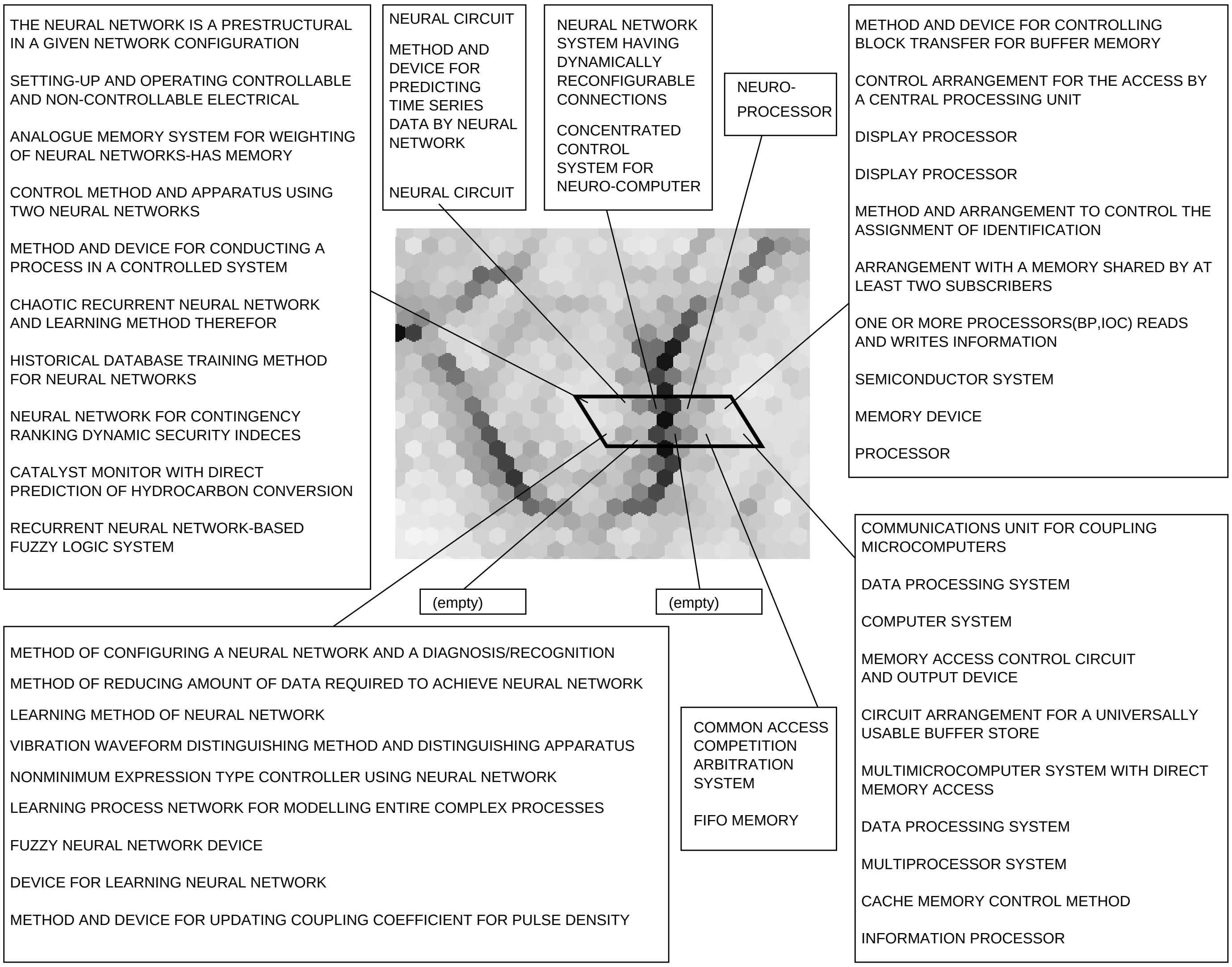

Fig. 2. Titles of the patent abstracts mapped near the border of the neural cluster detected with the novel method. The image in the center has been enlarged from Figure 1. The dark stripes represent abrupt changes in the density gradient. The titles of the abstracts picked from both sides of the dark stripe show how the topics of the patent abstracts change when moving across the stripe.

class called "Computer systems based on biological models" which includes neural networks. The method used for the visualization of clusters was able to discover a novel meaningful cluster within these data.

We also studied the patent abstracts in the areas where the visualizations of the U-matrix and the new cluster visualization method were more different. One of these areas is magnified in Figure 3, and randomly chosen titles have been shown for nodes in a sample area. The titles reveal that right below the dark stripe in the image there is a uniform cluster of patent abstracts about document processing, whereas in the area right above the dark stripe the topics of the patent abstracts are not focused clearly on any specific topic. This change in the topic is not reflected in the visualization of the U-matrix which has a coarser resolution. Thus, the new cluster visualization method was able to find a meaningful cluster which could not be perceived by the U-matrix alone.

The cluster visualization method was additionally able to find some of the borders of the patent classes of the old patent classification system. This is demonstrated in Figure 4, where the classes are seen to be located mostly on separate areas of the map grid. Note that this is only indirect evidence in favor of the method—all patent classes are not expected to form non-overlapping clusters.

V. Interpretation of the SOM structure as a set of local factors

In classical factor analysis, the factors are usually interpreted in terms of the "factor loadings," the correlations between the factors and the variables. In this work we will consider only one variant of the factor analysis framework, the principal component analysis (PCA), and we will not consider the subtleties distinguishing PCA from the common factor analysis and its variants. In PCA the factors are often called components, and the factor loadings are called component loadings, respectively. The data vector $\mathbf{x}$ is approximated by $\hat{\mathbf{x}}$ which is a linear combination of a

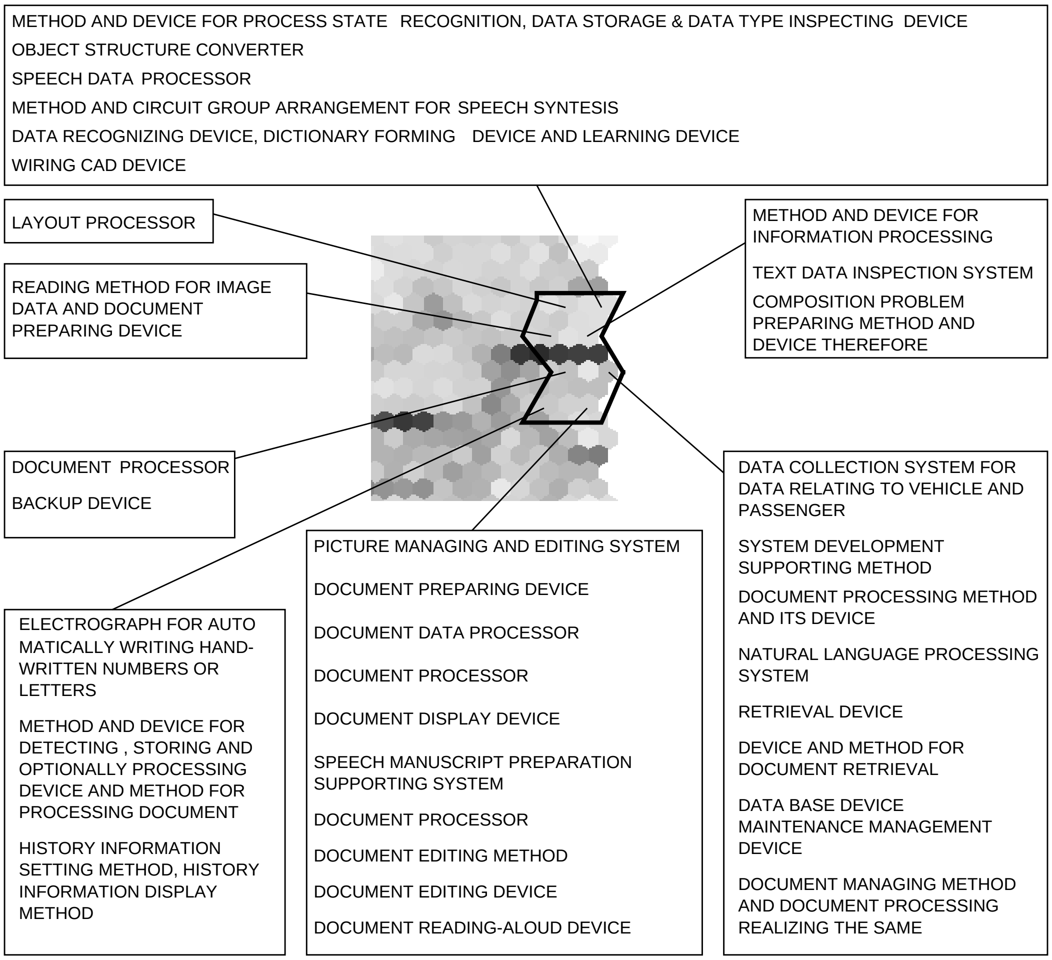

Fig. 3. The titles of patent abstracts around a gap between clusters. Above the dark stripe in the image the patents have (seemingly) miscellaneous titles, whereas below the dark stripe the titles focus strongly on document processing. The dark stripe therefore describes the sharp change in the density structure of the data.

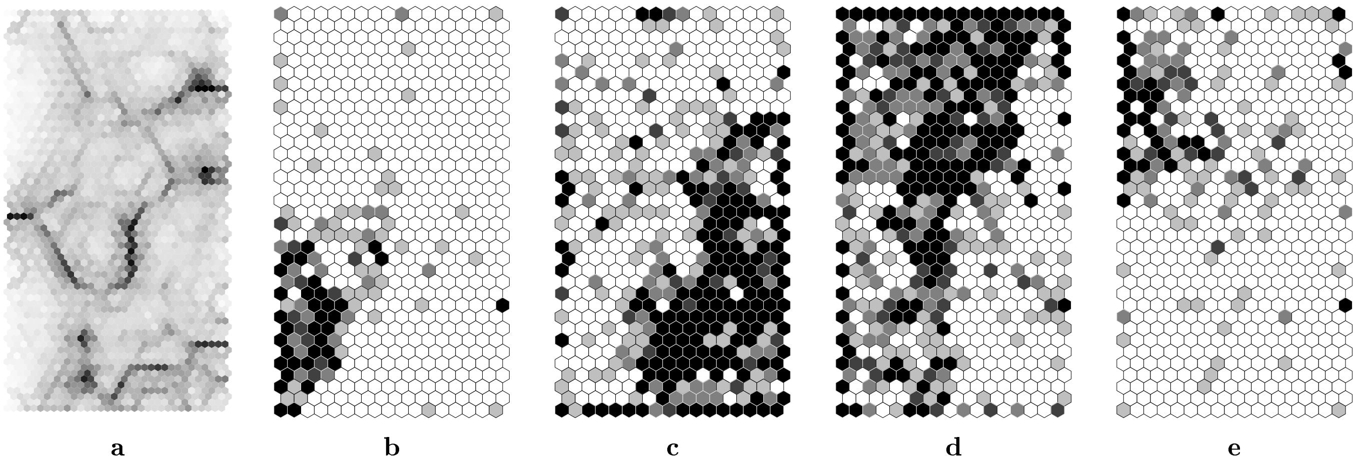

Fig. 4. The cluster visualization compared to the distribution of the original patent classes on the map grid. **a** Many of the dark stripes in the cluster visualization correspond to the class borders of the four known patent classes (b, c, d, and e). **b** Optical computing devices. **c** Electrical digital data processing. **d** Analogue computers. **e** Hybrid arrangements. The gray shade in b-e denotes number of patents of the given class in each map unit (white = 0, black = 4 or more).

set of factors,

$$\hat{\mathbf{x}} = \mathbf{A}\mathbf{f} \,. \tag{8}$$

Here $\mathbf{f}$ is a vector consisting of the factors, and both $\mathbf{f}$ and the matrix $\mathbf{A}$ are estimated from data so that as much of the variance of the data $\mathbf{x}$ is explained by the reconstructions $\hat{\mathbf{x}}$ as possible. It is well known that the optimal solution for a fixed number of factors is to choose the columns of $\mathbf{A}$ to be the eigenvectors of the covariance matrix of the data that correspond to the largest eigenvalues. The factors or principal components are then the projections of the data on the same eigenvectors.

It can be shown that in PCA the loading of factor j for variable i is equal to A_{ij}, the ith component of the jth eigenvector. In nonlinear maps however, the "loadings" may be different in different subdomains of the data, and they need to be measured locally at each point of the map. But how can one define a "local data set?" An additional problem is that if the data is high-dimensional the total number of loadings is very large: the dimensionality times the size of the map times the number of factors.

Our approximate solution is to assume that the SOM has already found a meaningful fit, and the "local factors" are then computed from the neighboring model vectors that describe the local orientation of the data distribution in a robust way. It may sound as unorthodox that the original data are not used, but there has not existed any other possibility to define a "local data set" than to use the locally smoothed values that relate to each other in a mutually compatible way. We also introduce a new method for summarizing the contributions using graphical overviews on the SOM groundwork. The method is especially useful for characterizing cluster structures in terms of variables that have the largest discriminating power between neighboring clusters.

A. *Local factors and contributions*

We have discussed in Sec. II how the local direction of the nonlinear two-dimensional elastic network formed by the SOM algorithm may be regarded to approximate the direction of local principal variation in the data. In this section we will characterize the direction using local factors and the contributions of the variables on them.

If the elastic network were a continuous, smooth nonlinear surface, it could be approximated locally by its linear tangent plane, and within the tangent plane it would be possible to apply locally linear interpretation methods. However, since the elastic network is formed of discrete elements and is therefore not smooth, the tangent plane must be approximated. Two basis vectors, fitted to the model vectors within a map neighborhood with the PCA algorithm, will span the plane. We shall define the two basis vectors as being the two *local factors* of the SOM.

Interpretation of the two local factors individually is somewhat arbitrary because of the classical problem in factor analysis of finding the "rotation" of the factors. Fortunately, however, we are not even interested in analyzing the factors separately at the moment since we are looking for methods for summarizing the loadings. If we consider the two factors together there is no arbitrariness. In linear factor analysis the contribution of a variable to the analysis may be summarized by the proportion of the variance of the variable explained by the factors together. The proportion is called the *communality*, and it can be computed as the sum of squares of all factor loadings for the variable. The communalities of the variables will be used to measure the contribution on the nonlinear map as well, but then the communalities need be computed locally using the two local factors.

Note 1. The interpretation of the communalities in terms of the explained variance of the data is not as straightforward for the nonlinear model of the data as it was for the PCA and factor analysis. That is why we will call the communalities computed for local factors *contributions.* The communalities have, however, a geometric interpretation as well, and the generalization of the geometric interpretation is more straightforward. The eigenvectors corresponding to the factors span a subspace in the data space. Denote the unit vector directed along the ith axis by $\mathbf{u}_i$. The communality of the ith variable is then equal to the squared length of the projection of $\mathbf{u}_i$ to the subspace. The squared length of the projection is a sensible measure of the "contribution" of the unit vector to the direction of the subspace. In the nonlinear models the contribution is measured in precisely the same fashion, but the global subspace is replaced by the local tangent plane of the nonlinear manifold.

Note 2. In traditional factor analysis the communality has often interpretations like reliability of the variable. The value is computed in the same fashion in our analysis but the interpretation is slightly different.

B. *Overview of the contributions of the variables*

The contribution of one of the original data variables to the organization of the map can be visualized by plotting the contribution as gray scales on the map groundwork. The display is light in areas where the contribution is large and dark where it is weak.

The problem with such contribution displays is, however, that there will be n such displays if the dimensionality of the data is n. In the text document map of Figure 1 there would be 988 displays in total! Summaries of relevant loadings are clearly needed. We have suggested earlier [12] that a cluster, for example, could be characterized by a summary indicating variables that have the largest contribution in the area surrounding it. Such variables have the largest discriminatory power between the cluster and the neighbor clusters.

The earlier summarization methods still had one shortcoming: The variables that best characterize a cluster in terms of the discriminatory power vary from neighboring cluster to cluster, and may even be different for different fine structures of the clusters. It was hard to take this variability into account with the earlier methods.

We will next introduce a method that visualizes how similar the contributions in neighboring map units are. Areas in which the the contributions of all variables are similar can then be grouped together when characterizing the clusters.

The similarity of the contributions can be compared by collecting the contributions of all variables to a vector, a "contribution profile," and comparing the profiles. The dissimilarity of the contribution profiles at neighboring map units can then be measured by the norm of the difference of the vectors, and visualized on top of the map groundwork. The difference between the profiles of neighboring units is converted to gray shades and plotted in between the units. In the resulting displays light shades will correspond to domains in which the profiles are similar and dark shades to domains in which the variables having the largest discriminatory power change.

C. *Case study: Characterization of a patent cluster*

We characterized the cluster of neurocomputing patents found with the new cluster visualization method described in Sec. IV-D. In order to find the variables (here words) discriminating the cluster from the neighboring clusters we studied the contributions of the variables to the local factors around the cluster. Due to the nonlinearity of the map and the data, the importance of the various variables is different in different regions surrounding the cluster. This was taken into account by dividing the border area into four relatively homogeneous regions according to the similarity of the contribution profiles (cf. Fig. 5). The contribution profiles of each region were then summarized by one average contribution profile, and the ten most strongly contributing words were chosen from amongst the total of 988 words.

The chosen words characterized the difference between the cluster and the area on the other side of the border. For example, to the right from the cluster of the neurocomputing patents there are mainly patent abstracts treating basic processes and parts of computers. This is reflected in the most contributing words of the two rightmost regions in Fig. 5, where for example the words "signal," "neural," "address," "access," "learn," and "memory" can be found.

The most contributing words in different border regions also reflected differences between the neighboring areas. In Figure 5 the words printed in capital letters are characteristic of only one or two regions, while the other words describe the overall difference between the neural cluster and the surrounding areas. For instance, in the border regions on the left the most contributing words include "optical," "Fourier," and "transform" which cannot be found in the border regions on the right. These words are typical to the patents treating optics and image processing. This observation is confirmed by considering the class distributions in Fig. 4. To the left from the cluster there are patents from the class of optical computing devices, whereas to the right there is a class related to general electrical digital data processing.

VI. Conclusions

We have introduced new SOM-based methods for the exploratory analysis of the cluster structures of high-dimensional data sets, and applied the methods to the analysis of a collection of patent abstracts. The SOM constructs an overview of the metric relations within the data set, and the overview was complemented by a new method that visualizes the clusteredness in the data. More specifically, the method detects clear between-cluster domains; they are signs of cluster borders or other interesting structures in data which need not consist of nicely-shaped clusters.

The cluster structures were interpreted using local factors that describe the local principal variation of the data. The contributions of the original variables in the factors can be summarized, for example in domains surrounding a detected cluster, and the summaries can be used for characterizing the cluster.

Acknowledgements

We thank Ms Krista Lagus for her help in constructing the WEBSOM user interface that we used while interpreting the patent collection.

References

[1] Teuvo Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, pp. 59–69, 1982.

[2] Teuvo Kohonen, *Self-Organizing Maps*, Springer, Berlin, 1995, (Second, extended edition 1997).

[3] Samuel Kaski, Jari Kangas, and Teuvo Kohonen, "Bibliography of self-organizing map (SOM) papers: 1981–1997," *Neural Computing Surveys*, vol. 1, no. 3&4, pp. 1–176, 1998, Available in electronic form at http://www.icsi.berkeley.edu/~jagota/NCS/: Vol 1, pp. 102–350.

[4] Trevor Hastie and Werner Stuetzle, "Principal curves," *Journal of the American Statistical Association*, vol. 84, pp. 502–516, 1989.

[5] Michael LeBlanc and Robert Tibshirani, "Adaptive principal surfaces," *Journal of the American Statistical Association*, vol. 89, pp. 53–64, 1994.

[6] Filip Mulier and Vladimir Cherkassky, "Self-organization as an iterative kernel smoothing process," *Neural Computation*, vol. 7, pp. 1165–1177, 1995.

[7] Helge Ritter, Thomas Martinetz, and Klaus Schulten, *Neural Computation and Self-Organizing Maps: An Introduction*, Addison-Wesley, Reading, MA, 1992.

[8] Timo Honkela, Samuel Kaski, Krista Lagus, and Teuvo Kohonen, "Newsgroup exploration with WEBSOM method and browsing interface," Tech. Rep. A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 1996.

[9] Teuvo Kohonen, Samuel Kaski, Krista Lagus, Jarkko Salojärvi, Jukka Honkela, Vesa Paatero, and Antti Saarela, "Self organization of a massive document collection," *IEEE Transactions on Neural Networks*, Accepted for publication.

[10] A. Ultsch, "Self-organizing neural networks for visualization and classification," in *Information and Classification*, O. Opitz, B. Lausen, and R. Klar, Eds., pp. 307–313. Springer-Verlag, Berlin, 1993.

[11] Samuel Kaski, Jarkko Venna, and Teuvo Kohonen, "Coloring that reveals high-dimensional structures in data," in *Proceedings of ICONIP'99, 6th International Conference on Neural Information Processing*, T. Gedeon, P. Wong, S. Halgamuge, N. Kasabov, D. Nauck, and K. Fukushima, Eds., vol. II, pp. 729–734. IEEE Service Center, Piscataway, NJ, 1999.

[12] Samuel Kaski, Janne Nikkilä, and Teuvo Kohonen, "Methods for interpreting a self-organized map in data analysis," in *Proceedings of ESANN'98, 6th European Symposium on Artificial Neural Networks, Bruges, April 22-24*, Michel Verleysen, Ed., pp. 185–190. D-Facto, Brussels, Belgium, 1998.

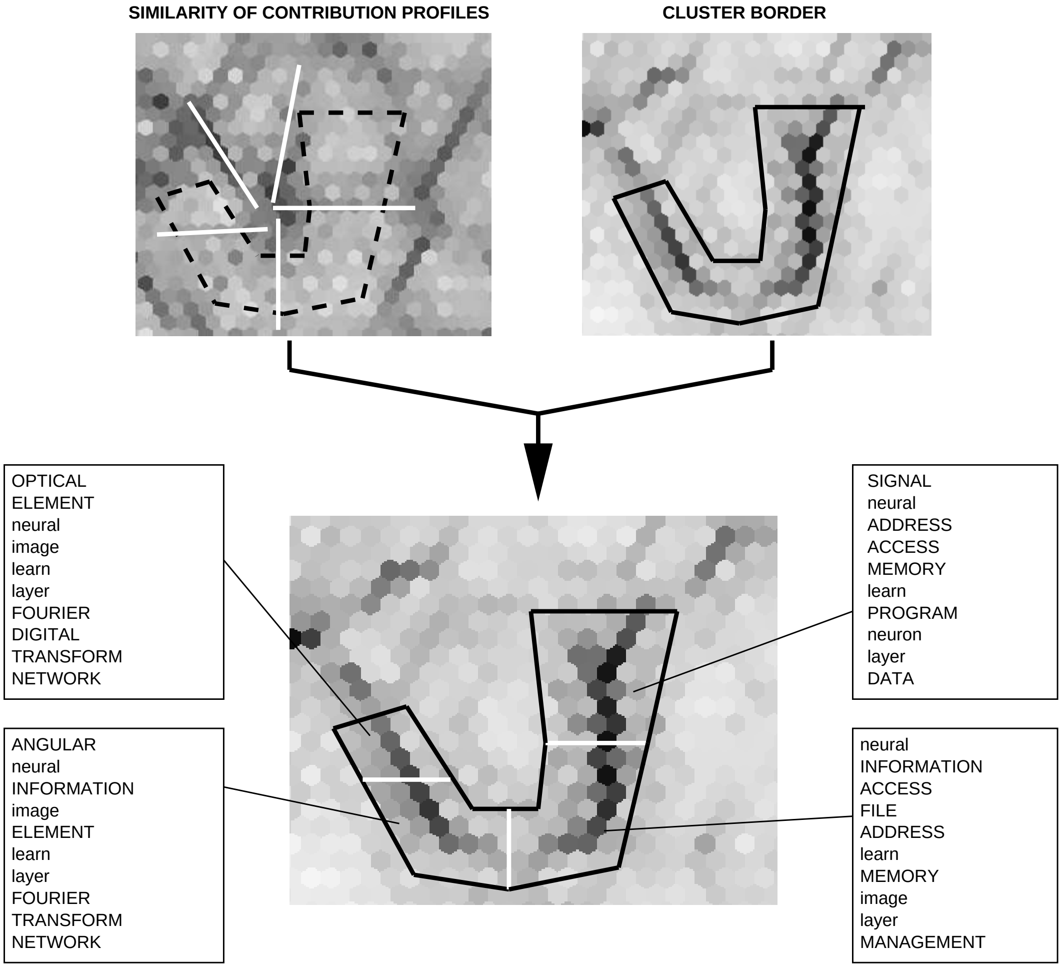

Fig. 5. The most important variables (words) contributing to the difference between the cluster of neurocomputing patents and the surrounding map areas. The image on the upper left is the summary of the contribution profiles; on light areas the contribution profiles of neighboring nodes are similar and on dark areas dissimilar, respectively. Areas where the contribution profiles are relatively homogeneous have been separated with the white lines, and combined with the detected cluster border shown on the upper right. The end result is the border line of the neural cluster divided into four regions. The ten most contributing words have been highlighted for each region. The words characteristic to only one or two regions have been printed with capital letters.