Proc. ICONIP'99, to appear

Coloring that Reveals High-Dimensional Structures in Data

Samuel Kaski, Jarkko Venna, and Teuvo Kohonen

Helsinki University of Technology
Neural Networks Research Centre
P.O. Box 5400, FIN-02015 HUT, Finland
samuel.kaski@hut.fi

Abstract

We introduce a method for assigning colors to displays of cluster structures of high-dimensional data, such that the perceptual differences of the colors reflect the distances in the original data space as faithfully as possible. The cluster structure is first discovered with the Self-Organizing Map (SOM), and then a new nonlinear projection method is applied to map the cluster structure into the CIELab color space. The projection method preserves best the local data distances that are the most important ones, while the global order is still discernible from the colors, too. This allows the method to conform flexibly to the available color space. The output space of the projection need not necessarily be the color space, however. Projections onto, say, two dimensions can be visualized as well.

1. Introduction

In exploratory data analysis or interactive data mining one central goal is to illustrate high-dimensional data sets by overviews that are easily understandable but still preserve the essential properties of the data. The Self-Organizing Map (SOM) algorithm (Kohonen, 1982; Kohonen, 1995) can be used for visualizing one of the most central properties of high-dimensional statistical data, namely, its cluster structure, on graphical map displays.

In this paper we introduce a method for visualizing the cluster structures discovered by the SOM with colors. Unlike in commonly used visualizations of data types or clusters the (relative) colors are not chosen arbitrarily, but the *perceptual differences of the colors are supposed to reflect the distance relations within the cluster structure* as faithfully as possible.

This work was supported by the Academy of Finland.

The perceptual differences between different colors can be approximated by distances in the color space called CIELab (CIE, 1986). Therefore, any coloring of high-dimensional data items actually corresponds to a mapping of the items into a smaller-dimensional space, the CIELab color space. In this work we do not map the data items themselves but their representations, the models formed by the SOM algorithm. A nonlinear projection method that is suitable for the SOM models will be introduced in Section 3, and applied to coloring cluster structures in Section 4.

2. Discovering Cluster Structures with the Self-Organizing Map

The SOM algorithm can be used to form two-dimensional graphical displays that are visual overviews of data sets. The SOM is a regular grid onto which the data is mapped in an orderly fashion: neighboring grid points represent similar data and the mapping therefore converts the original high-dimensional relations in the data into simpler, visualizable two-dimensional relations.

The mapping is defined by attaching a model vector $\mathbf{m}_i \in \mathbb{R}^n$ to each point of the regular SOM grid. Here n is the dimensionality of the data. Each data item is mapped to the location on the map that has the closest model vector. As a result of the SOM algorithm, the model vectors become ordered and, additionally, the density of the model vectors reflects the density of the data in the input space.

The ordered SOM grid can be used as a groundwork for displaying the cluster structure of the data. Clusteredness, i.e. the density of the data, is reflected in the density of the model vectors, which in turn is reflected in the distances between neighboring model vectors. This idea has been utilized in the U-matrix method (Ultsch, 1993) in which the distances are visualized by gray levels. Figure 1 illustrates the cluster structure of a data set consisting of 39 statistical indicators that describe various aspects of poverty of each country in the world (Kaski and Kohonen, 1996).

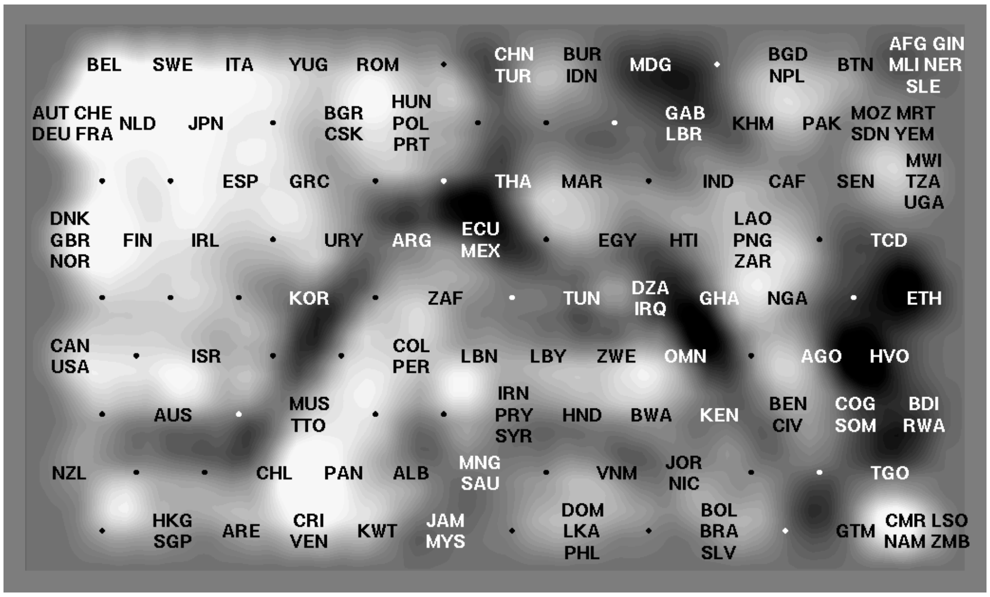

Figure 1. Structured diagram of the data set (World Bank, 1992) chosen to describe the standard of living. The order of the abbreviated country names indicates the similarity of the standard of living of the countries, and the gray levels, obtained by smoothing the U-matrix, indicate the degree of clustering. Light shades correspond to a high degree of clustering and dark shades to gaps in the degree of clustering. Different types of welfare and poverty are visible as the clustered (light) areas on the map, separated by the dark "ravines". The SOM grid points having no countries have been marked with dots.

3. SOM-Based Non-Linear Projection that Preserves Local Distances

When projecting a set of high-dimensional data vectors into a lower-dimensional space it is in general impossible to preserve *all* of the pairwise distances between the data items. In the mapping of Sammon (cf. the end of this section), a compromise is made by minimizing an objective function that describes the *global error*, errors in all distances.

Fortunately, preservation of all distances is not equally important. When aiming at a projection that preserves the cluster structures, i.e. the density of the data, it is especially important to preserve the *local distances* that are indicators of the local density. In the SOM the model vectors are already ordered so that the distances between model vectors of neighboring map units are the local distances. This is the reason why distances between neighbors are used in the SOM-based clustering displays to indicate the clusteredness. We introduce below a new nonlinear projection method that aims at preserving these local distances.

Denote the index set corresponding to the neighborhood of map unit i by N_i (here excluding the map unit i itself), and denote the distance between the model vectors i and j by $d_{ij} = \|\mathbf{m}_i - \mathbf{m}_j\|$. If the low-dimensional projection of $\mathbf{m}_i$ is denoted by $\mathbf{m}'_i$, and the distance after the projection by $d'_{ij} = \|\mathbf{m}'_i - \mathbf{m}'_j\|$, then

$$E_1 = \sum_i \sum_{j \in N_i} (d_{ij} - d'_{ij})^2 \qquad (1)$$

is the mean-square cost function that measures distortions in the local distances. The projection method should find such values for the $\mathbf{m}'_i$ that the cost function E_1 is minimized.

Unfortunately, a projection that preserves the local distances is not guaranteed to be globally ordered. The projection may "fold" over itself and become unintelligible.

It has turned out in our experiments that the projection becomes globally ordered if an only slightly restrictive extra constraint is imposed on the non-local distances: Intuitively speaking, no other model vector may be projected in between the model vectors belonging to neighboring points on the SOM grid. The detailed definition is given and illustrated in Figure 2. The exact value of the non-local distances may be arbitrary as long as they are large enough.

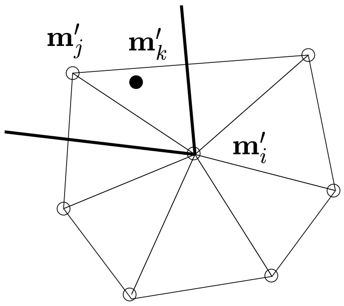

Figure 2. Definition of orderliness of the projection. The projections of the model vectors in the neighborhood of the unit i have been depicted with open circles, and neighbors have been connected with thin lines. The most obvious definition of orderliness might be that the map may not become "twisted", i.e., no non-neighbor may enter the area bordered by the outer thin lines. This definition has disadvantages, however: it is not easily applicable, especially in higher-dimensional spaces, and not as obvious as here if the neighborhood is very irregular. We have therefore used a more straightforward definition: If a non-neighbor $\mathbf{m}'_k$ belongs to the "sector" of $\mathbf{m}'_j$, illustrated with the thick lines, and is closer to $\mathbf{m}'_i$ than $\mathbf{m}'_j$ is, then $\mathbf{m}'_k$ is too close to $\mathbf{m}'_i$. The sector consists of those points for which the angle from $\mathbf{m}'_j$, as seen from $\mathbf{m}_i$, is smaller than from the other neighbors of the middle point $\mathbf{m}'_i$.

The orderliness can be guaranteed by adding a second term into the cost function (1). Denote by I_{ij} the set of indices of all such model vectors that do not satisfy the orderliness constraint for the pair of neighbors (i, j). The total squared deviation from orderliness can then be measured with

$$E_2 = \sum_{i,j} \sum_{k \in I_{ij}} (d'_{ij} - d'_{ik})^2 \ . \qquad (2)$$

The total cost function to be minimized is then

$$E' = E_1 + \lambda_2 E_2 \ . \qquad (3)$$

Here λ_2 is a parameter that determines how strongly deviations from orderliness affect the projection. We recommend initializing the projection in an orderly fashion to avoid local minima, and setting λ_2 to a relatively large value, say, $\lambda_2 = 100$.

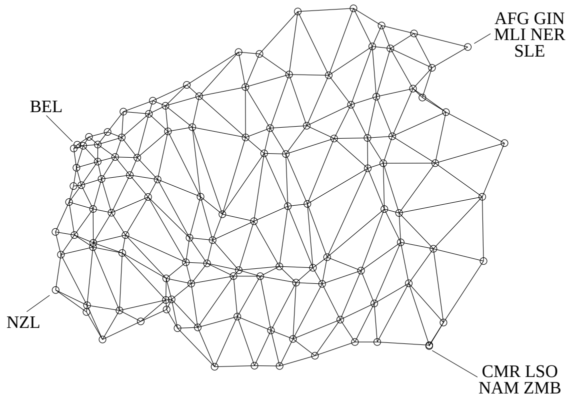

Figure 3. Projection of the model vectors of the SOM of Figure 1 onto the two-dimensional space. The small circles denote the projections of the model vectors. Projections of model vectors of neighboring map units have been connected with lines, and countries mapped to the corner units have been marked in the figure.

The projection of the SOM of Figure 1 into a two-dimensional space, obtained by minimizing E' in (3), has been shown in Figure 3. We optimized the cost function by a simple stochastic gradient descent: At each optimization step one model vector was selected randomly and moved into the direction of the negative gradient of E' by an amount that decreased gradually during the optimization. If the gradient step would have increased the value of the cost function, the step size was temporarily halved.

It can be seen from Figure 3 that the long distances between the projected model vectors correspond to black areas in Figure 1 and short distances to light areas, respectively.

Relation to Sammon's Mapping. The traditional multidimensional scaling methods like the Sammon's mapping (Sammon, Jr., 1969) could in principle be used instead of the method presented here. They do not, however, produce as flexible mappings as our method, since they try to represent *all* of the pairwise distances *of the original data.* The local differences will then necessarily be represented less accurately in Sammon's mapping, except in special cases. Moreover, if only a restricted area of the output space is available like in the coloring task discussed in Section 4 below, then it may be more difficult to fit Sammon's projection to the area because the mapping is "stiffer".

Note: Sammon's mapping could be applied to the model vectors of the SOM as well. The cost function of Sammon's mapping would then be

$$E_S = \sum_{i \neq j} \frac{1}{d_{ij}} (d_{ij} - d'_{ij})^2 \ ,$$

(omitting a constant normalizing term). Sammon's mapping would consider distances between *all* model vector pairs (i, j), weighted by $1/d_{ij}$, whereas for the present purpose we consider only the distances between model vectors of neighboring SOM units. Therefore, Sammon's mapping would still be less flexible.

4. Coloring by Projection into the Color Space

Coloring corresponds to projection into a color space. We have used the CIELab space (CIE, 1986) which has been defined so that perceptual differences in colors correspond to Euclidean distances in the space as well as possible, at least locally.

The whole three-dimensional space is not available for the coloring, however. A given display device like the CRT tube is able to present only a part of the space. We have restricted the available space even further since we wanted the colors to differ predominantly in one perceptual quality only, namely, the hue. Therefore, very non-saturated colors (more exactly, colors with a small chroma) were not allowed, and the lightness was set to a fixed value. After these restrictions a non-regularly shaped area in a two-dimensional cross section of the CIELab space remained (cf. Fig. 4).

We need an additional term in the cost function of the projection to keep the projected items within the available region. Before formulating the term we need to *scale* the distances of the CIELab space suitably, however, since the relative scale between the distances in the original data space and in the CIELab space may be arbitrary. A proper choice of the scaling factor, denoted by κ, will be discussed in more detail later.

Denote, as previously, the ith model vector of the SOM by $\mathbf{m}_i$ and its projection by $\mathbf{m}'_i$, respectively. Denote then the point in the available space that is closest to $\mathbf{m}'_i/\kappa$ by $\bar{\mathbf{m}}'_i$. If $\mathbf{m}'_i/\kappa$ is already in the available space, then $\bar{\mathbf{m}}'_i = \mathbf{m}'_i/\kappa$. The additional term, E_3, in the cost function is then the squared distance between the actual projection and the closest available point, scaled to match the distances in the original data space,

$$E_3 = \sum_i \|\kappa \bar{\mathbf{m}}'_i - \mathbf{m}'_i\|^2 \ . \qquad (4)$$

The total cost function that the projection shall minimize is

$$E = E_1 + \lambda_2 E_2 + \lambda_3 E_3 \ , \qquad (5)$$

where λ_2 and λ_3 are suitably chosen parameters. We have minimized E by the stochastic gradient descent method described in Section 3. The cost function has many local minima, and thus the projection should be initialized properly, in an orderly fashion. A suitable sample initialization is shown in Figure 4**a**.

The parameter λ_2 should be set to a large value, say, $\lambda_2 = 100$ like in Section 3. We recommend setting the other parameter, λ_3, initially to zero to make the mapping flexible, and to increase its value gradually to, say, $\lambda_3 = 10$ during the optimization to reinforce the constraint.

The scaling factor κ in (4) is the only remaining free parameter in the cost function. Since the region of the CIELab space that corresponds to the available colors is very limited and irregularly shaped, a compromise must be made between a faithful representation of the distances and the color resolution. The scale should be chosen according to their relative importance.

If $1/\kappa$ is small, like in Figure 4**b**, the distances are represented faithfully but only a small part of the available color space is utilized. As $1/\kappa$ increases (Figs. 4**c** and **d**) the projection utilizes a larger portion of the available space but the distances between model vectors become somewhat more distorted because the projection must conform to the borders of the area.

Fortunately, however, small changes in the scale usually only have small effects on the coloring. For example, there are only small distortions in the colorings corresponding to either of the projections in Figs. 4**c** and **d** although the projections seem very different (see `http://www.cis.hut.fi/~sami/iconip99/`).

A suitable initial estimate for κ might be obtained by first scaling the largest distance in the input space to the maximal distance in the color space, and then choosing some proportion of the obtained scale.

The result of mapping the model vectors of the SOM of Figure 1 into the CIELab color space is shown in Figure 5. Compared with the non-constrained projection in Figure 3 the overall organization is similar but some compromises have clearly been needed especially in conforming to the disk of non-saturated colors in the middle. Nevertheless, it is clear, based on the result, that the projection method is able to make the necessary compromises.

When the graphical SOM display was colored according to the projection (see `http://www.cis.hut.fi/~sami/iconip99/`), the perceptual differences in the hue of the neighboring map units corresponded well with the distances in the original data space, depicted as shades of gray in Figure 1. In addition, the hues became ordered globally; inside the different clustered areas the hues were relatively uniform, and different clustered areas attained different colors.

5. Coloring of the Original Data According to Its Cluster Structure

The coloring can be even more useful if the original data set can be visualized in some other manner besides the SOM display. Then each data item can be colored according to its color on the SOM display. The welfare and poverty structures, for instance, can be visualized in a straightforward manner on a geographic map display. The countries can then be colored according to their welfare or poverty type (see `http://www.cis.hut.fi/~sami/iconip99/`).

The result is a display where countries having a similar welfare or poverty type have been

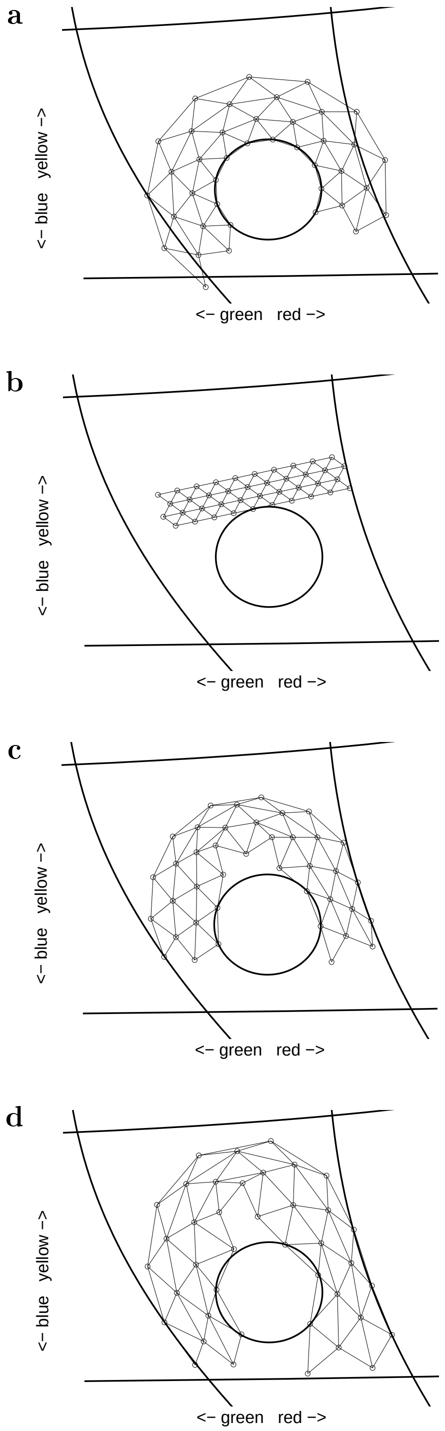

Figure 4. Effect of scale on the projection of an artificially generated, evenly spaced two-dimensional SOM into a constant-lightness cross section of the CIELab color space. **a**: Initial state. **b**: Optimal preservation of distances but low resolution: small scale $1/\kappa$. **c** and **d**: Progressively larger scale. The thick lines delimit the region representable by a typical CRT tube, and the circle in the middle encompasses colors that were considered too non-saturated (gray).

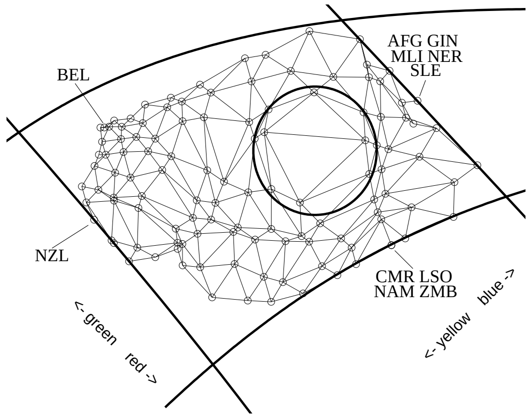

Figure 5. The projection of the model vectors of the SOM of Figure 1 into a constant-lightness cross section of the CIELab color space. The small circles denote the projections of the model vectors. Projections of model vectors of neighboring map units have been connected with thin lines. The thick lines delimit the region representable by a typical CRT tube, and the circle in the middle encompasses colors that were considered too non-saturated (gray). Note that the axes are different from those used in Figure 4.

colored similarly irrespective of their geographical location. Japan and Australia, for example, are fairly similar to the European countries and to the USA and Canada. Countries that belong to very different types than their neighbors pop out strongly, like Japan, Sri Lanka, and South Africa.

Such visualizations can be very useful if the data has a "natural" ordering like the geographical order here but, in fact, *any* order of the data items can be used. For example, if the countries were ordered simply according to the GNP per capita in a table and colored using a SOM, then countries in which the welfare or poverty type is different from the other countries having a similar value of GNP per capita would be clearly discernible based on sharp discontinuities in the coloring.

Note: In principle our method could be used to map the data set directly, instead of mapping the model vectors of a SOM computed from the data set. It would, however, be more difficult to define which distances are local enough so that they should be represented accurately. The SOM makes this decision automatically. If it is necessary to obtain a characteristic color for each data item then some suitable local fitting method may be used to complement the SOM-based mapping.

6. Conclusions

We have introduced an automatic method for coloring graphical SOM displays so that the perceptual properties of the coloring reflect the properties of the high-dimensional statistical data as closely as possible. In fact, the resulting coloring is almost tailored to the human color vision system which is very accurate in detecting differences between the colors of neighboring areas, in this case the color differences between neighboring locations on cluster displays. The easily interpretable coloring makes it possible to visualize complex statistical structures automatically even for non-experts.

References

CIE (1986). Colorimetry, 2nd ed. CIE Publication No. 15.2.

Kaski, S. and Kohonen, T. (1996). Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world. In Refenes, A.-P. N., Abu-Mostafa, Y., Moody, J., and Weigend, A., editors, *Neural Networks in Financial Engineering*, pages 498–507. World Scientific, Singapore.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69.

Kohonen, T. (1995). *Self-Organizing Maps.* Springer, Berlin. (Second, extended edition 1997).

Sammon, Jr., J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18:401–409.

Ultsch, A. (1993). Self-organizing neural networks for visualization and classification. In Opitz, O., Lausen, B., and Klar, R., editors, *Information and Classification*, pages 307–313. Springer-Verlag, Berlin.

World Bank (1992). *World Development Report 1992.* Oxford University Press, New York.