11 Statistical Data Analysis by the Self-Organizing Map

Samuel Kaski and Teuvo Kohonen

Knowledge discovery in databases (KDD) [1], sometimes also referred to as data mining, is a recently established field of research in which the aim is to discover novel patterns or structures in large data sets. The complex interactive discovery process involves several stages, of which the central stage called data mining refers to the application of essentially any suitable methods for finding interesting patterns in data.

KDD is related to a field of statistics called exploratory data analysis. Statistical inferences are often made in a two-stage process. Hypotheses are first generated in a data-driven phase, and the hypotheses are tested in another, confirmatory phase. The first methods for the exploratory, data driven phase were developed already in 1970's. The increase in computing power allows us to use much more sophisticated methods for looking at the statistical structures in data, and to analyze much larger data sets.

The central goal in exploratory data analysis is to present a data set in a form that is easily understandable but at the same time preserves as much essential information of the original data set as possible. The exploratory data analysis methods are general-purpose instruments that illustrate the essential features of a data set, like its clustering structure and the relations between its data items.

One may distinguish two categories of exploratory data analysis tools with somewhat different goals. First, some tools like the Sammon projection [5] *project* the multidimensional data set to, e.g., a two-dimensional plane while trying to preserve its whole structure (the distances between the data items) as well as possible. Other methods [3] try to find *clusters* in the data, whereby instead of the large data set only a small number of clusters needs to be considered.

A vast number of different algorithms to perform clustering is available. Choosing suitable algorithms and applying them correctly requires thorough knowledge of both the algorithms and the data set. There must exist enough clustering tendency in the data set in order that the use of clustering algorithms would be sensible at all, and as different clustering algorithms tend to find clusters of different shapes, the suitability of the shapes to describe the data set must be verified.

The projection methods, on the other hand, do not reduce the amount of data to be presented. Although they illustrate the essential features of the data set, the illustration is costly to obtain and may still be difficult to understand if the data set is large.

The self-organizing map algorithm is a unique method in that it combines the goals of both the projection and the clustering algorithms. It can be used at the same time to visualize the clusters in a data set, and to represent the set on a two-dimensional map in a manner that preserves the nonlinear relations of the data items; nearby items are located close to each other on the map. Moreover, even if no explicit clusters exist in the data set, the self-organizing mapping method reveals "ridges" and "ravines". The former are open zones with irregular shapes and high clustering tendency, whereas the latter separate data subsets that have a different statistical nature.

11.1 Case Study: Structures of Welfare and Poverty in the World

In this study we have demonstrated how the Self-Organizing Map is able to describe structures in a macroeconomic system. The map is shown to illustrate the "welfare or poverty states" of the countries of the world, when the data set describes different aspects of the standard of living. State transitions can easily be followed on the map. It is hoped that this study would serve as a recipe on how, using standard procedures, the state of any micro- or macroeconomic system can be presented in an easily understandable form. Only the data set needs to be changed in different applications.
The understanding and description of a complex entity like the standard of living requires simultaneous consideration of a large collection of statistical indicators describing its different aspects and their relationships. In this study we used a total of 39 indicators that described factors like health, education, consumption, and social services, picked up from the World Development Report of the year 1992 [7]. Based on the set of statistical indicators the SOM can be used to represent the welfare and poverty "states" of the countries on a *"poverty map"*.

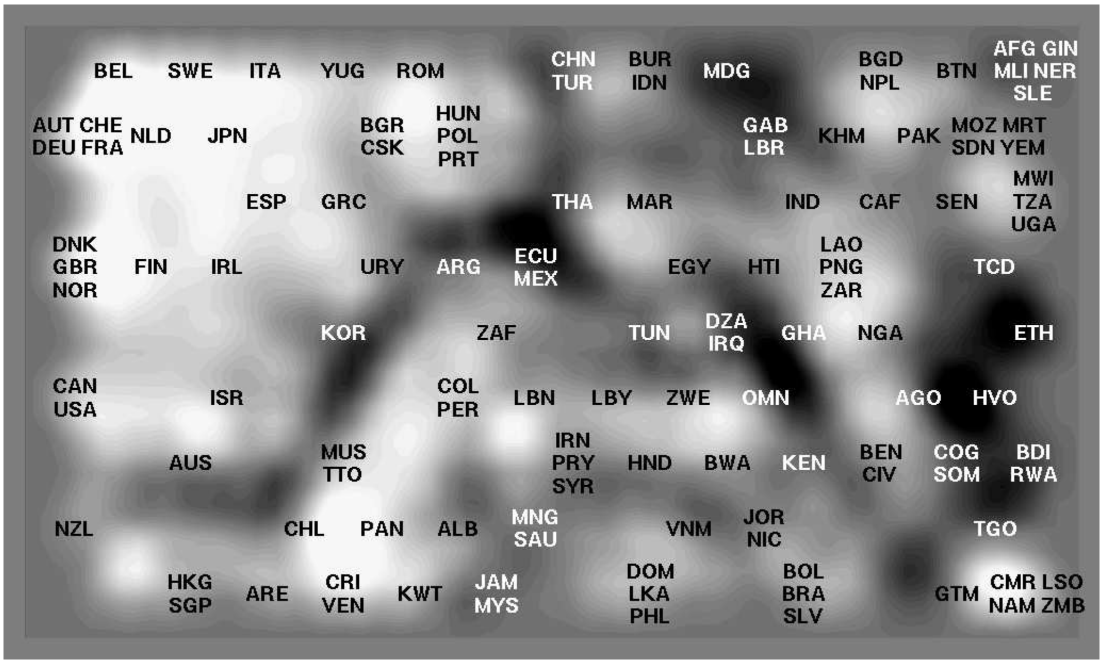

Figure 13: Structured diagram of the data set chosen to describe the standard of living. The order of the abbreviated country names indicates the similarity of the standard of living of the countries, and the colors indicate the degree of clustering. Light areas represent areas of a high degree of clustering and dark areas 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".

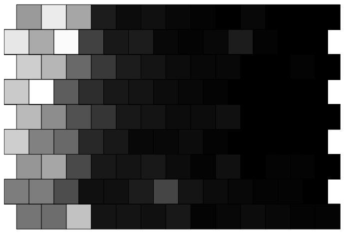

Figure 14: Distribution of the GNP per capita, which was not used in computing the maps, shown over the SOM groundwork. White indicates the largest value in the material and black the smallest, respectively. The horizontal axis of the map seems to correlate with the distribution of the overall welfare, as measured by the GNP per capita.

The clustering tendency in the data set can be visualized as a false-color or gray-scale display on the map (Figure 13; the display is a smoothed version of the so-called U-matrix [6] display; cf. also [2]). Different *types of welfare and poverty* are manifested on the display as clustered areas. For example, the cluster in the top left corner consists predominantly of the OECD countries, and an annex on the right of it is a cluster consisting mostly of the countries of Eastern Europe. It is evident from the display that most of the clustered areas are neither regularly shaped nor easily separated, but instead form some kinds of “hills”, “ridges” and “ravines”. It is then a definitive advantage of the method that no assumptions need to be made about the cluster shapes before the analysis, as is implicitly done in many other clustering methods.
The overall order of the countries on the map was found to illustrate the traditional conception of welfare; in fact, the horizontal dimension of the map seems to correlate fairly closely with the GNP (Gross National Product) per capita (Figure 14).
Refined interpretations about the fine structure of the welfare and poverty types, the clustered areas on the map display, can be made based on the original statistical indicators. The values of the indicators can be displayed in their natural order on the groundwork formed of the organized map. This display is much more easily understandable that ordinary linear statistical tables (examples have been shown in Fig. 15). Furthermore, such displays are readily amenable to interactive exploration using suitable computer interfaces. For example, a click on an interesting location on the cluster display (Figure 13) might highlight the corresponding location on the indicator displays of Figure 15.

11.2 Case Study: Country Risk Ratings

Similar displays can be created to illustrate any data set. Another economic data set that may be of interest to economists is a set of nine indicators published by the Euromoney magazine (March 1996): Country risk ratings. In the display shown in Figure 16, like in the previous display of welfare and poverty, Finland is situated in the top left corner, together with, e.g., most West-European countries.
These kinds of displays provide an overview of the state of the world at a given moment, and a possibility to explore the state further. It would of course be of interest also to follow the changes in the state for several years. The Self-Organizing

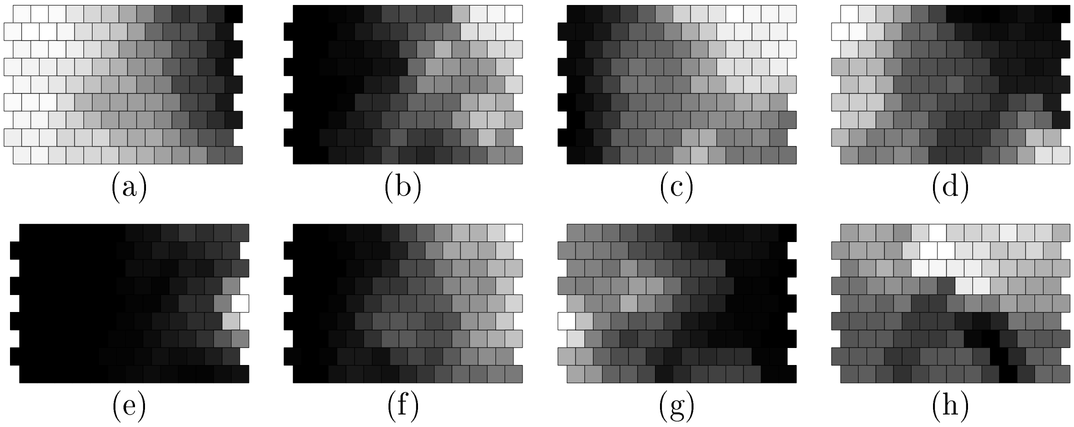

Figure 15: The values of some of the indicators visualized on the SOM groundwork. Since the countries have been organized into a natural order, the displays have clear "patterned" outlook instead of being purely random. Therefore they can be interpreted quickly. (a) Life expectancy at birth (years); (b) Adult illiteracy (%); (c) Share of food in household consumption (%); (d) Share of medical care in household consumption (%); (e) Population per physician; (f) Infant mortality rate (per thousand live births); (g) Tertiary education enrollment (% of age group); and (h) Share of the lowest-earning 20 percent in the total household income. In each display, white indicates the largest value and black the smallest, respectively.

Map is readily amenable also for such studies.

11.3 Conclusions

The SOM has been applied to case studies to show how it can be used as a decision-support system, to get a quick but yet quite accurate impression of the structures inherent in any data set.
Following exactly the same procedures the SOM could also be used for analyzing and visualizing sets of statistical indicators in other similar applications. For instance, the method has already been used for the analysis of states of banks [4]. The SOM formed a "solvency map," from which the state of the banks could be inferred at a glance. In time series analysis it is important that the nature of change in the state of the banks can be visualized on the map (e.g., as a slow shift toward the bankrupt region) even if the changes could not be predicted by more traditional methods.

References

[1] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 1–34. AAAI Press / MIT Press, Menlo Park, CA, 1996.

[2] J. Iivarinen, T. Kohonen, J. Kangas, and S. Kaski. Visualizing the clusters on the self-organizing map. In C. Carlsson, T. Järvi, and T. Reponen, editors, *Proceedings of the Conference on Artificial Intelligence Research in Finland*,

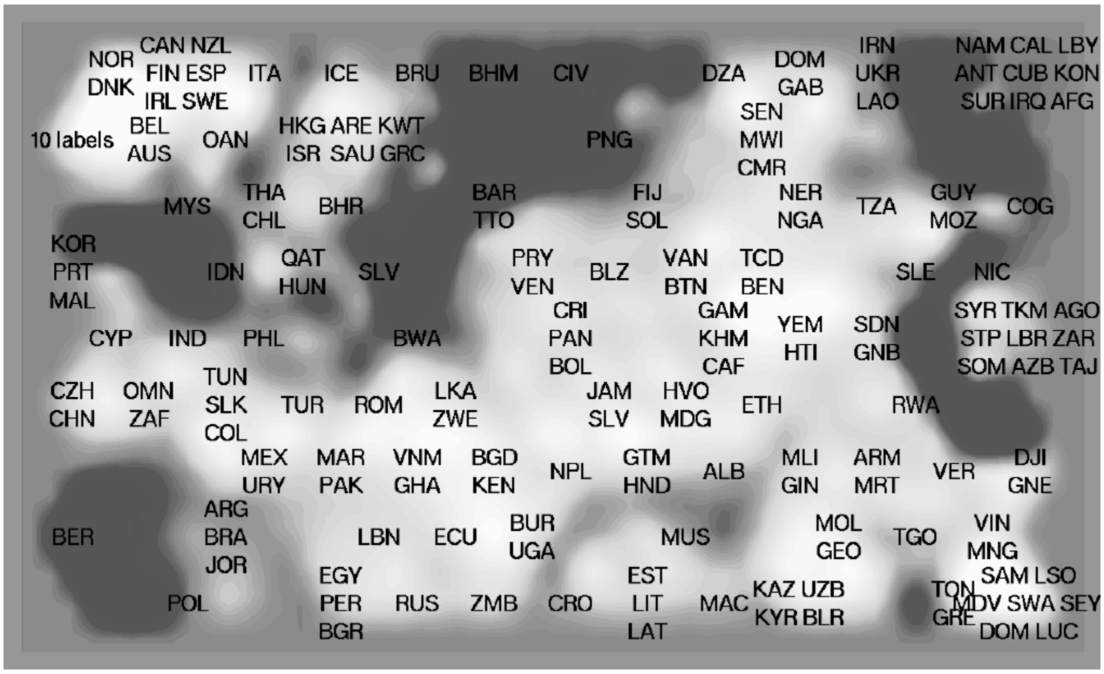

Figure 16: Country risk ratings (Euromoney, March 1996). An illustration of the structures within a data set that describes different aspects of the country risk (economic performance, political risk, debt, access to finance etc.). Based on the distribution of the original indicators on the map groundwork (not shown) it may be summarized that, e.g., the countries in the upper right hand corner have significant amounts of debts, and countries in the lower right hand corner perform poorly economically. Countries in the upper left hand corner perform best on every indicator. (The "10 labels" in the image refers to countries LUX, CHE, SGP, JPN, USA, NLD, DEU, AUT, GBR, and FRA.)

number 12 in Publications of the Finnish Artificial Intelligence Society, pages 122–126. Finnish Artificial Intelligence Society, Helsinki, Finland, 1994.

[3] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice Hall, Englewood Cliffs, NJ, 1988.

[4] B. Martín-del-Brío and C. Serrano-Cinca. Self-organizing neural networks for the analysis and representation of data: Some financial cases. *Neural Computing & Applications*, 1:193–206, 1993.

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

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

[7] World Bank. *World Development Report 1992*. Oxford Univ. Press, New York, NY, 1992.