Chapter 9

Self-organizing map

Teuvo Kohonen, Merja Oja, Samuel Kaski, Panu Somervuo

9.1 Self-Organizing Maps: introduction

Teuvo Kohonen

The name Self-Organizing Map (SOM) signifies a class of neural-network algorithms in the unsupervised-learning category. In its original form the SOM was invented by the founder of the Neural Networks Research Centre, Professor Teuvo Kohonen in 1981-82, and numerous versions, generalizations, accelerated learning schemes, and applications of the SOM have been developed since then.

The central property of the SOM is that it forms a nonlinear projection of a high-dimensional data manifold on a regular, low-dimensional (usually 2D) grid. In the display, the clustering of the data space as well as the metric-topological relations of the data items are clearly visible. If the data items are vectors, the components of which are variables with a definite meaning such as the descriptors of statistical data, or measurements that describe a process, the SOM grid can be used as a groundwork on which each of the variables can be displayed separately using grey-level or pseudocolor coding. This kind of combined display has been found very useful for the understanding of the mutual dependencies between the variables, as well as of the structures of the data set.

The SOM has spread into numerous fields of science and technology as an analysis method. We have compiled a list of over 4000 scientific articles that apply the SOM or otherwise benefit from it.

The most promising fields of application of the SOM seem to be

- data mining at large, in particular visualization of statistical data and document collections,
- process analysis, diagnostics, monitoring, and control,
- biomedical applications, including diagnostic methods and data analysis in bioinformatics, and
- data analysis in commerce, industry, macroeconomics, and finance.

References

[1] Teuvo Kohonen, *Self-Organizing Maps*, Springer Series in Information Sciences, Vol. 30, Springer, Berlin, Heidelberg, New York, 1995, 1997, 2001, 3rd edition.

9.2 4465 Works on SOM

Merja Oja, Samuel Kaski, Teuvo Kohonen

The Self-Organizing Map (SOM) algorithm has attracted a great deal of interest among researches and practitioners in a wide variety of fields. The SOM has been analyzed extensively, a number of variants have been developed and, perhaps most notably, it has been applied extensively within fields ranging from engineering sciences to medicine, biology, and economics. We have collected a comprehensive list of 4465 scientific papers that use the algorithms, have benefited from them, or contain analyses of them. The list is intended to serve as a source for literature surveys.

The collection is available at the WWW address `http://www.cis.hut.fi/nnrc/refs/` (cf. [1]). Additions to the list and error reports are most welcome; please send any correspondence to the email address `biblio@mail.cis.hut.fi`.

Evolution of SOM applications. We studied how SOM research within certain application areas has evolved during the years. A set of 13 topical categories was selected by combining classes used by the INSPEC (tm) database. The number of articles in those categories was then plotted as a function of the year of publication. The plots only contain the articles available in the INSPEC collection.

The results shown in Figure 9.1 reveal, for instance, that speech recognition was a very popular topic already in the beginning of 90's, while there still are applications in that area. Some other disciplines, including "information science and documentation" and "business and administration," are still gaining popularity.

A SOM of SOM references. The SOM references were organized onto a document map to study the relationships between the topic categories, and to provide an interface for browsing and searching the collection. A WEBSOM [2] was computed using the titles of the documents. For some documents also an abstract was available and it was used in the computation.

The map is available for browsing and search in the address `http://websom.hut.fi/websom/somref/search.cgi`.

References

[1] Samuel Kaski, Jari Kangas, and Teuvo Kohonen. Bibliography of self-organizing map (SOM) papers: 1981–1997. *Neural Computing Surveys*, 1(3&4):1–176, 1998. Available in electronic form at http://www.icsi.berkeley.edu/∼jagota/NCS/: Vol 1, pp. 102–350.

[2] 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*, 11:574–585, 2000.

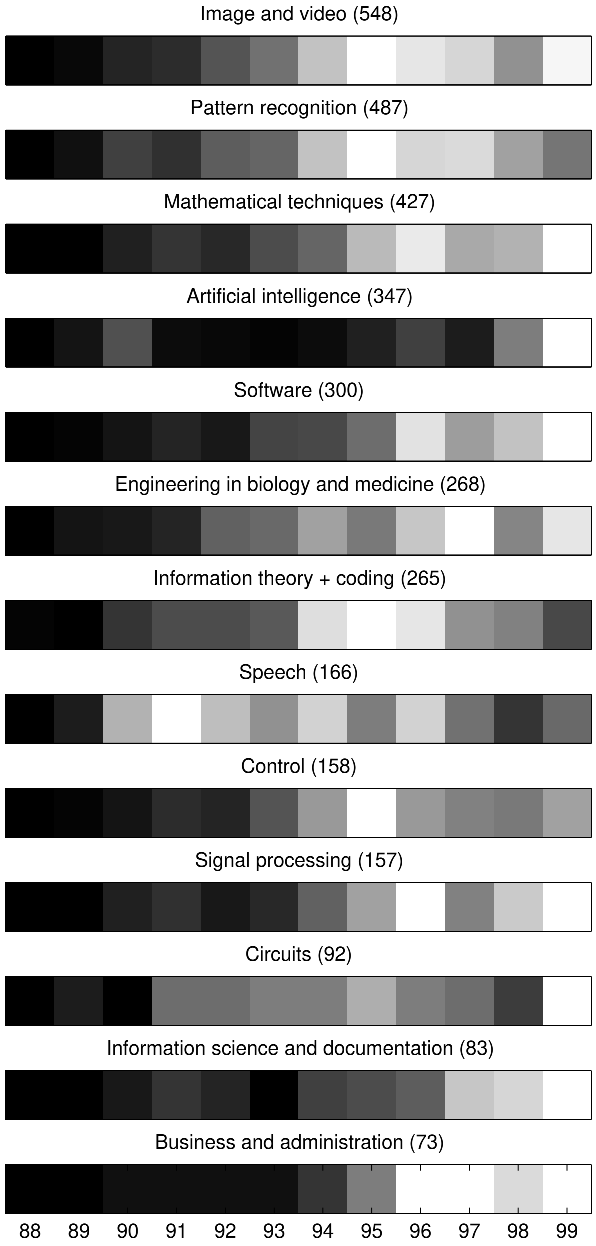

Figure 9.1: Evolution of SOM applications. Each horizontal bar shows the number of published SOM works within a certain topic category, as a function of the publication year shown in the bottom. The total number of works in the category is shown in parentheses. White: Largest number within the category, black: relatively small number of articles.

9.3 Clustering and visualization of large protein sequence databases by means of an extension of the self-organizing map

Panu Somervuo and Teuvo Kohonen

The amount of DNA sequences, protein sequences, and molecule structures studied and reported, e.g., in the Internet is already overwhelming. One should develop better tools for the analysis of the existing databases. Among the new challenges one may mention finding the hidden relations between the data items, revealing structures from large databases, and representing the results to the human in a comprehensible way. The classification and clustering of the sequences may reveal new unknown connections between them. The visualization of large data sets in a compact way may give insights into the data and lead to the development of new ideas and theories.

The Self-Organizing Map of symbol strings [2] was used in this work [4] for the clustering of all the 77,977 protein sequences of the SWISS-PROT database, release 37 [1]. In this method, unlike in some previous ones, the data sequences are not converted into histogram vectors in order to perform the clustering. Instead, a collection of true representative model sequences that approximate the contents of the database in a compact way is found automatically, based on the concept of the generalized median of symbol strings, after the user has defined any proper similarity measure for the sequences. The FASTA method [3] was used in this work. The benefits of the SOM and also those of its extension are fast computation, approximate representation of the large database by means of a much smaller, fixed number of model sequences, and an easy interpretation of the clustering by means of visualization. The complete sequence database is mapped onto a two-dimensional graphic SOM display, and clusters of similar sequences are then found and made visible by indicating the degree of similarity of the adjacent model sequences by shades of gray. The geometrically organized picture makes it possible to illustrate the relationships of a large amount of sequences at a glance, see Fig. 9.2.

Besides the protein sequences, we have applied the extension of the SOM also to the clustering of the protein molecules based on their three-dimensional molecule structures, see Fig. 9.3.

References

[1] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.*, 27:49-54, 1999.

[2] T. Kohonen. Self-organizing maps of symbol strings. Report A 42, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 1996.

[3] W. Pearson and D. Lipman. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 85:2444-2448, 1988.

[4] P. Somervuo and T. Kohonen. Clustering and Visualization of Large Protein Sequence Databases by Means of an Extension of the Self-Organizing Map. 3rd International Conference on Discovery Science, Kyoto, Japan, Dec. 4-6, 2000, *Lecture Notes in Artificial Intelligence* 1967, pages 76–85, 2000.

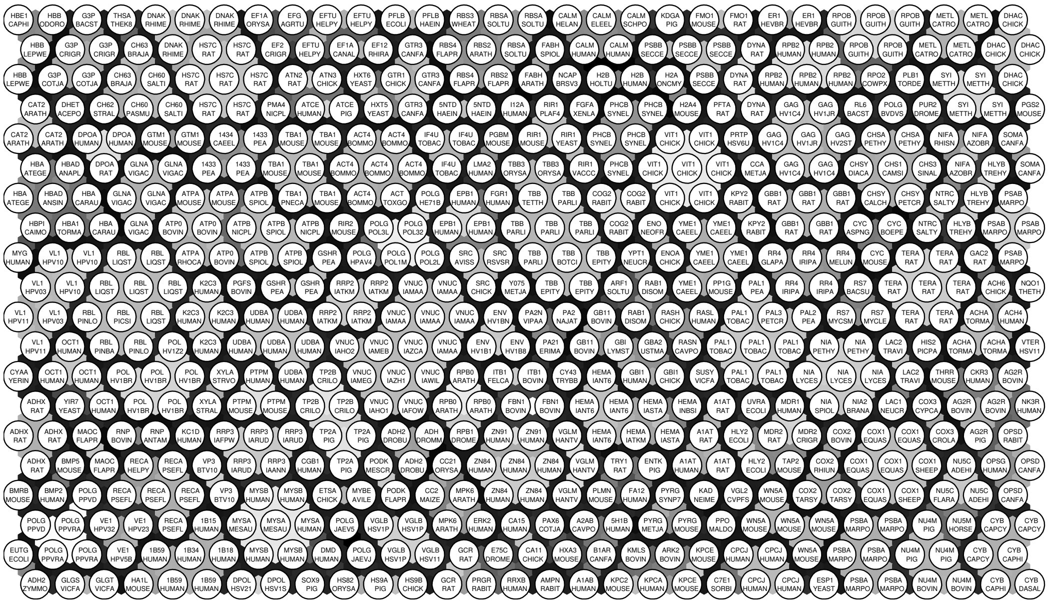

Figure 9.2: A 30-by-20-unit hexagonal SOM grid. The SOM was constructed using all the 77,977 protein sequences of the SWISS-PROT release 37. Each node contains a prototype sequence and a list of data sequences. The labels on the map nodes are the SWISS-PROT identifiers of the prototype sequences. The upper label in each map node is the mnemonic of the protein name and the lower label is the mnemonic of the species name. The similarities of the neighboring prototype sequences on the map are indicated by shades of gray. The light shades indicate a high degree of similarity, and the dark shades a low degree of similarity, respectively. Light areas on the map reveal large clusters of similar sequences.

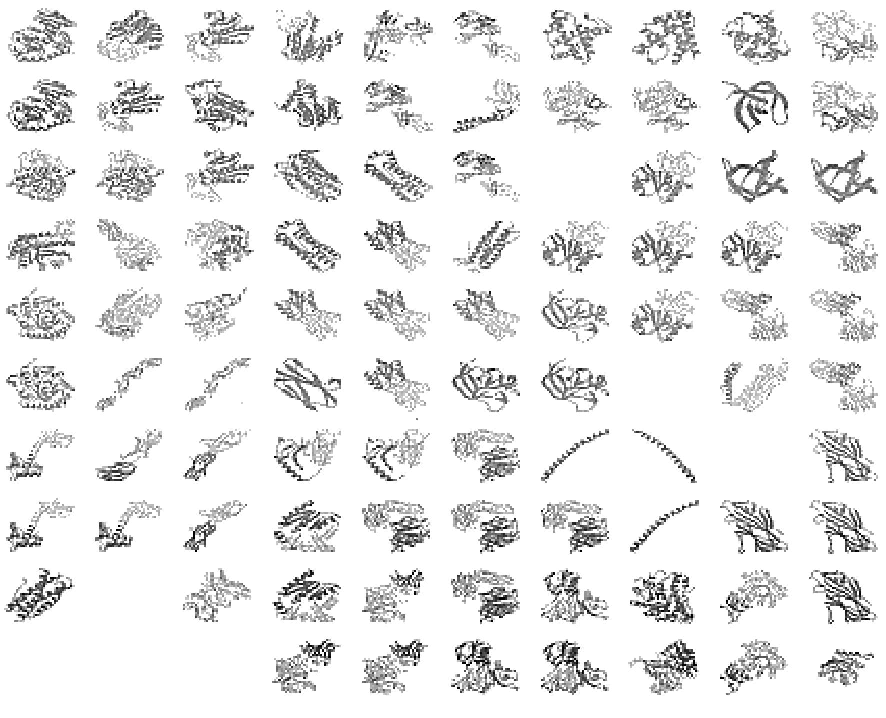

Figure 9.3: The SOM of 3D protein molecules.