CMA4CH 2010, Mediterraneum Meeting

Application of Multivariate Analysis and Chemometry to Cultural Heritage and Environment
3rd ed., Taormina, Sicily island, Italy, Europe, 26-29 September 2010

Self-Organizing Maps: Learning for Exploratory and Predictive Modelling

R.G. Brereton
Centre for Chemometrics, School of Chemistry, University of Bristol, Cantocks Close, Bristol BS8 1TS, UK

Abstract

Self-organizing maps (SOMs) involve using machine learning methods to visualize different patterns in data and to determine the relationship between experimental measurements and samples.

SOMs were first reported by the Finnish professor Teuvo Kohonen, and is sometimes called a Kohonen map [1,2]. For 30 years the SOMs have been widely employed for visualization of relationships between samples.

Unsupervised SOMs as traditionally employed are used primarily for exploratory data analysis to reveal relationships between samples in data. They allow visualization of a large number of samples in limited space.

However, it is also possible to employ SOMs for classification purposes whereby an additional vector of class information is included in the training.

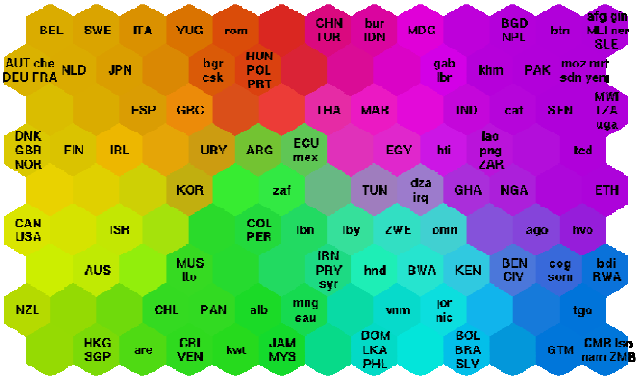
Introduction

SOMs, as originally described, can be considered as an unsupervised method, which is an alternative to principal components analysis (PCA) since they can be used to present the characteristic structure of data in a low-dimensional space. However, they also can be modified for using in a supervised mode. SOMs are neural networks that employ adaptive learning algorithms, so are well-suited for many real systems. A SOM involves a map, which is often represented by a two-dimensional grid although can be represented in spherical or circular representations, or any desired geometry.

SOMs are a powerful alternative to PCA for several reasons. First, PCA models are linear but in many complex situations we expect nonlinearities in the data; second, there are many facile ways of graphical display using SOMs, and third, PC models can be strongly influenced by outliers, common problems especially with complex data sets as occur in biology and cultural heritage studies.

These approaches are much more computationally intensive compared to the traditional statistical methods and so have not been used a lot by analytical chemists in the past. However, at the time of writing with rapid increase in computing power, these calculations can be performed in real time using modern scientific desktop computers. Probably because they are computationally intensive there is a limited packaged software available.

Data analysis, clustering and visualization by the SOM can be done using either public domain, commercial, or self-coded software. In Bristol a comprehensive set of Matlab routines have been developed for display of SOMs.

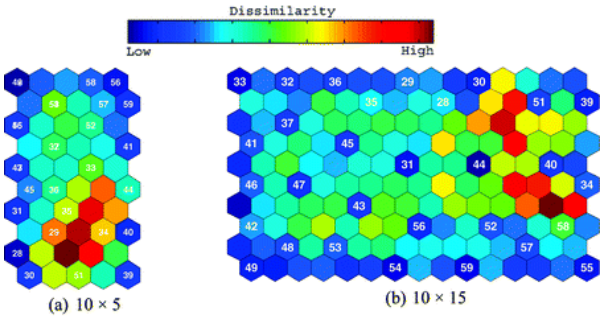


Examples

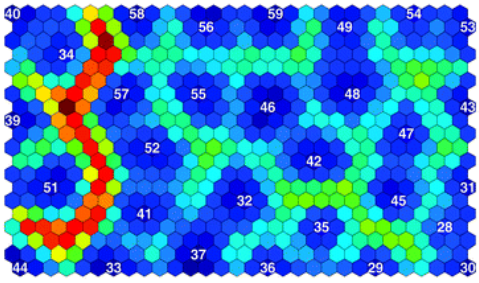
Today there are a large number of application of SOMs in many fields. The use of SOMs is not restricted to chemistry but one of the most famous examples of an unsupervised SOM is the World Poverty Map with the analysis and visualization of large collections of statistical and macroeconomic data, produced in 1999 using data of 1992 by Helsinki University, Laboratory of

Computer and Information Science, see figure on previous page. Another example, using a supervised SOM, by this author, is an application in the area of metabolic profiling, consisting of a nuclear magnetic resonance (NMR) data set of 96 samples of human saliva, characterized by three factors [3], namely, whether the sample has been treated using a mouthwash or not, 16 donors, and 3 sampling days, differing for each donor. The SOMs can be supervised using each of the different factors. A supervised SOM for the donors using optimal scaling values are shown in the figure on the left. It can be seen that there is a separation between groups on these maps for all cases, especially for the minor factors. Note that for many donors, the samples now fall into one region of the map rather than two regions, see on the right. A key to this approach is that the variables associated with each factor can be visualised using component planes.

Another example is the use of SOMs in quality control, allowing the development of a map of the normal operating condition (NOC) samples, applied to online monitoring in the case of high-performance liquid chromatography (HPLC) measurements of a continuous pharmaceutical process [4]. The figures on the right are the U-matrix visualizations for the trained map consisting of the 30 NOC samples where the numbers represent the BMU of each sample. The distance of a future sample to the map can be determined, the greater this distance the more likely the sample is an outlier.



(a) 10 × 5 (b) 10 × 15


(c) 20 ×30

Conclusions

Unlike traditional methods such as PLS or PCA, PLS-DA or LDA, SOMs are more rarely employed by chemometricians for pattern recognition. SOMs, although common in many areas of This is probably because the software is not so widely available in packaged form aimed at analytical chemists, and is more computationally intensive than traditional algorithms. It is possible however to develop quantitative measures based on SOMs, to determine for example which variables (e.g. peaks in GCMS) are best as discriminators and whether an unknown sample fits into a predefined group, and hence adapt existing approaches to situations encountered in chemometrics.

References

1) T. Kohonen, Construction of Similarity Diagrams for Phonemes by a Self-Organising Algorithm; Helsinki University of Technology: Espoo, Finland, 1981.
2) Kohonen, T., Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43:59-69 (1982)
3) K. Wongravee, G.R. Lloyd, C.J. Silwood, M. Grootveld, R.G. Brereton, Supervised Self Organizing Maps for Classification and Determination of Potentially Discriminatory Variables: Illustrated by Application to Nuclear Magnetic Resonance Metabolomic Profiling, *Anal. Chem.*, 82(2), (2010) 628–638
4) S. Kittiwachana, D.L.S. Ferreira, L.A. Fido, D.R. Thompson, R.E.A. Escott, R.G. Brereton, Self-Organizing Map Quality Control Index, *Anal. Chem.*, 82(14), (2010) 5972–5982