Analysing Dissimilarity Matrices via Kohonen Maps

Christophe Ambroise and Gérard Govaert

Université de technologie de Compiègne
URA CNRS 817
BP 649 F-60206 Compiègne cedex - France
Email: ambroise@hds.univ-compiegne.fr govaert@hds.univ-compiegne.fr

Summary: This paper introduces a method for analysing dissimilarity matrices. The proposed approach is based on the Kohonen self-organizing maps algorithm and may be used either for clustering or for data visualization. This is illustrated with the analysis of a real data set and compared with classical multidimensional scaling techniques.

1. Introduction

Exploratory data analysis aims to detect and analyse heterogeneity, variability, and underlying structures in important data sets. Most of the time, the data comprises a set of n objects described by several variables. This makes it, possible to define a distance on the space of the objects (also called input space) and then to use a great amount of existing methods which assume the existence of a distance (PCA, k-means...).

Sometimes the only available data is a dissimilarity matrix. Dissimilarity data are measures representing the amount of differences between pairs of objects. Dissimilarities among n objects are specified with an $n \times n$ matrix, $\delta = \{\delta_{ij}\}_{i,j=1..n}$. Dissimilarity matrices may arise from different origins: Human judgement may be easily translated into similarity or dissimilarity measurement. Data like driving times between pairs of cities are naturally available as dissimilarities and it is also possible to derive a dissimilarity matrix from an objects/variables data structure.

In some cases the transformation of the dissimilarity matrix into a distance matrix does not involve a great loss of information and classical methods may be used effectively. But the analysis of a dissimilarity matrix may require specific methods when the dissimilarities are badly approximated by distances. Multidimensional scaling (MDS) is a set of data analysis techniques that picture the structure of the set of objects. Multidimensional scaling techniques can be classified into metric and non metric approaches. The metric methods produce representation preserving quantitative information (Sammon 1969) and non metric methods concentrate on the qualitative relationship between the objects (Kruskal 1964).

An alternative class of methods which perform non–linear projection or feature extraction has been developed in the field of neural networks. A well known neural network algorithm is the Kohonen self-organizing maps algorithm (SOM, Kohonen 1982), which performs dimensionality reduction when the objets under consideration are described by a set of variables. We propose to adapt a variant of the Kohonen maps algorithm for dealing directly with dissimilarity data.

2. Kohonen Maps and dissimilarity data

On the one hand, the SOM algorithm is a clustering algorithm closely related to the k-means, on the other hand it is used for dimensionality reduction of high-dimensional data. The method associates a finite number of d-dimensional vectors, called input patterns, with a finite number of prototypes (also called neurons). The prototypes are organized in a one-dimensional or two-dimensional array. The topological relationship between input patterns in the input space is reflected as faithfully as possible in the arrangement of the corresponding neurons in the array (also called output space). In the neural networks terminology, this property is referred to as topology preservation. Used as a clustering technique, each neuron matches with a unique cluster and the relationship between the neurons may facilitate the interpretation of the partition. For dimensionality reduction, the input patterns are 'projected' onto the grid, and form a one or two-dimensional non-linear representation of the data.

The SOM method was originally implemented as an adaptive algorithm. Kohonen has presented a 'batch' version called the 'Batch Map' algorithm (Kohonen 1993). The authors have investigated the relationship between the probabilistic approach of clustering and the 'Batch Map' algorithm and have proposed a batch algorithm (Ambroise and Govaert, to appear) which may be easily adapted for dealing with dissimilarity data.

Let $\mathbf{c} = \{c_{kl}\}_{k,l=1..K}$ be the matrix of the distances between the K neurons in the ouptut space. Starting from a random classification of the objects, the following steps are computed until the classification becomes stable:

- **Finding** the object which is the best representative of each class. Object i^* is the prototype of class k if:

$$i^* = arg \min_i \sum_{j \in \mathcal{C}_k} \delta_{ij} \tag{1}$$

- **Topology preservation:** For each object i, K coefficients $(p_{i1}, \cdots, p_{iK})$ are computed:

$$p_{ik} = \frac{h(k, k^*)}{\sum_{k'=1}^{K} h(k', k^*)} \tag{2}$$

  where k^* is the class of object i, and $h(k, k^*)$ is the neighboring function,

$$h(k, k^*) = \begin{cases} 1 & \text{if } c_{kk^*} < \sigma \\ 0 & \text{otherwise} \end{cases}$$

  where σ is the width of the neighborhood taken into account and c_{kk^*} the distance between class k and class k^* in the output space. σ is decreasing with the number of iterations.

- **Stochastic classification:** Each object i is assigned randomly to a class k according to the multinomial distribution defined by the proportions $(p_{i1}, \cdots, p_{iK})$.

3. Numerical Example

3.1 The French rail way

The dissimilarity matrix that we have considered contains the travel times by train between 21 important French cities. This data has been pictured by three different methods: The Sammon projection, the Kruskal algorithm and a hybrid Kohonen–Sammon projection, described below.

3.2 Sammon Projection for Kohonen Maps

Classical representations of Kohonen maps display each object at the location of its prototype in the output space. Following Lowe and Tipping (1995), we propose to use Sammon mapping for generating a better display of a Kohonen map. The Sammon mapping finds $\mathbf{x}_i$,...,$\mathbf{x}_n$, a configuration of points in the plane, which minimizes the following stress function:

$$S = \frac{1}{\sum_{i>j} \delta_{ij}} \sum_{i>j} \frac{(\delta_{ij} - d_{ij})^2}{\delta_{ij}} \tag{3}$$

where d_{ij} is the euclidian distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. A hybrid representation between Kohonen maps and Sammon mapping can be obtained by replacing, in the stress function, the initial dissimiarities δ_{ij}, by δ^*_{ij}

$$\delta^*_{ij} = (1 - \alpha) \cdot \delta_{ij} + \alpha \cdot c_{class(i)class(j)} \tag{4}$$

where $c_{class(i)class(j)}$ is the distance between the prototypes of objects i and j in the output space.

3.3 Discussion

All the mappings show that the travel times are not directly linked with the geographical distances between the cities (Figure 1a).

The Sammon Mapping (Figure 1b) and the Kruskal MDS (Figure 1c) produce very similar representations. They locate *Paris* in a central position and tend to map the other cities around *Paris* in a circular structure, preserving the local relationship. This is indicative of the centralized structure of the French rail way.

The Kohonen like algorithm (Figure 1d), with 4 prototypes, separates the cities in four groups which correspond roughly to north, south/east, east and south/west geographical regions. The topology preserving property of this algorithm gives information about the similarity relationship existing between these four clusters. The Kohonen–Sammon projection ($\alpha = 0.8$) represents the cities while respecting the global information relative to the positioning of the clusters, and shows also the local relationship between the cities within each class.

In these figures, we have illustrated the efficiency of an algorithm for getting a Kohonen map from a dissimilarity matrix and the specificity of the Kohonen approach in dimensionality reduction. A Kohonen map projects the objects in a low dimensional space and produces a classification of these objects. The hybrid Kohonen–Sammon

(a) Geographical location of 21 french cities

(b) Sammon projection

(c) Kruskal projection

(d) Hybrid Kohonen–Sammon projection

Figure 1: The French rail way

projection, in its data representation, takes advantage of these two features for picturing the data.

References:

Ambroise, C. and Govaert, G. (To appear): Constrained Clustering and Kohonen Self-Organizing Maps. *Journal of Classification.*

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

Kohonen, T. (1993): Things you haven't heard about the self-organizing map. *Proceedings of 1993 IEEE International Conference on Neural Networks*, San Francisco, California, 1147–1156.

Kruskal, J.B. (1964): Non metric multidimensional scaling: A numerical method. *Psychometrika*, **29**, 2, 115–129.

Lowe, D. and Tipping M. (1995): A novel neural network technique for exploratory data analysis. Proceeding of ICANN1995, Paris, **1**, 434–440.

Sammon Jr, J.W (1969): A non linear mapping for data structure analysis. *IEEE Transactions on Computers*, **18**, 5, 401–409.