ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 216, 2008

*Andrzej Dudek**

KOHONEN SELF-ORGANIZING MAPS FOR SYMBOLIC OBJECTS

ABSTRACT. Visualizing data in the form of illustrative diagrams and searching, in these diagrams, for structures, clusters, trends, dependencies etc. is one of the main aims of multivariate statistical analysis. In the case of symbolic data (e.g. data in form of: single quantitative value, categorical values, intervals, multi-valued variables, multi-valued variables with weights), some well-known methods are provided by suitable 'symbolic' adaptations of classical methods such as principal component analysis or factor analysis. An alternative visualization of symbolic data is obtained by constructing a Kohonen map. Instead of displaying the individual items $k = 1, ..., n$ by n points or rectangles in a two dimensional space, the n items are first clustered into a number m of mini-clusters and then these mini-clusters are assigned to the vertices of a rectangular lattice of points in the plane such that 'similar' clusters are represented by neighbouring vertices in the lattice.

Article present algorithm of creating Kohonen self-organizing maps for symbolic objects along with some examples on datasets taken from symbolic data repository (http://www.ceremade.dauphine.fr/~touati/sodas-pagegarde.htm).

Key words: Classification, visualization, symbolic data, neural networks.

I. INTRODUCTION

Self-organizing maps (SOMs) are a data visualization technique invented by Professor Teuvo Kohonen which reduces the dimensions of data through the use of self-organizing neural networks. They can also be treated as a classification method due to assignments of mini-clusters of objects to nodes of rectangular lattice. Kohonen's algorithm has been developed for "traditional" numerical data. Recently some extensions of Kohonen's algorithm have been proposed by El Golli, Conan-Guez and Rossi [2004] adapting SOMs to data in form of intervals First part of this paper explains how to create self-organizing maps and locate SOMs among methods of multivariate statistical analysis. Second part is an.

* Ph.D., Chair of Econometrics and Informatics, University of Economics, Wrocław.

introduction to symbolic data analysis, symbolic objects and symbolic variables are described and dissimilarity measures for symbolic objects are presented. In third part some modifications of original Kohonen's algorithms extending SOMs onto data in form of intervals are described. Forth part presents examples of creation of SOM from data in form of intervals and use Kohonen map for symbolic objects as discriminant analysis technique. Finally some conclusions and remarks are given.

II. KOHONEN'S SELF-ORGANIZING MAPS AMONG OTHER TECHNIQUES OF MULTIVARIATE STATISTICAL ANALYSIS

Self-organizing maps are a data visualization technique but they also be can treated as classification method as well as a branch of neural networks. This method assumes than objects are first clustered into m mini-clusters and then these mini-clusters are assigned to the vertices of a rectangular lattice of points in the plane such that 'similar' clusters are represented by neighbouring vertices in the lattice.

The algorithm of self-organizing maps creation can be described in four main points (Kohonen[1997]).

1. Cluster prototypes (centers) are defined randomly or in some way (for example using eigenvectors of principal components).

2. Iteratively, input object (x) is assigned to cluster with the smallest distance to its prototype. As distance measure for this purpose squared euclidean distance is used (which is important for later considerations due to fact that euclidean distance is not defined for symbolic data).

3. Prototypes of clusters are re-calculated according to formula 1:

$$P_i \leftarrow P_i + \alpha^{(k)} h_{c(x),i}(x - P_i), \tag{1}$$

where:

P_i - cluster prototype; $\alpha^{(k)}$ - learning factor in k-th iteration step, typically this factor its decreasing according to k; $h(\cdot)$ – kernel neighbouring function, typically treshold, gaussian, Epanechnikov or exponential kernel; $c(x)$ – cluster, to which x is assigned.

4. Steps two and three are repeated until convergence criterion is fulfilled.

The final effect can be presented in form of mini-clusters map as in figure 1.

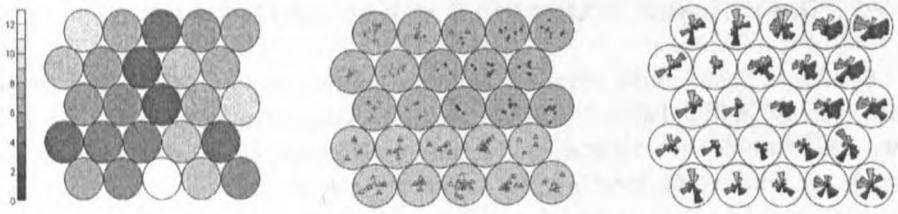

Figure 1. Self-organizing maps examples.
Source: Own research, graphics and calculations made in R environment with "Kohonen" library.

It's hard to assign SOMs exactly to one of techniques of Multivariate Statistical Analysis, but there are some relations between Kohonen's maps and other methods pointed in figure 2.

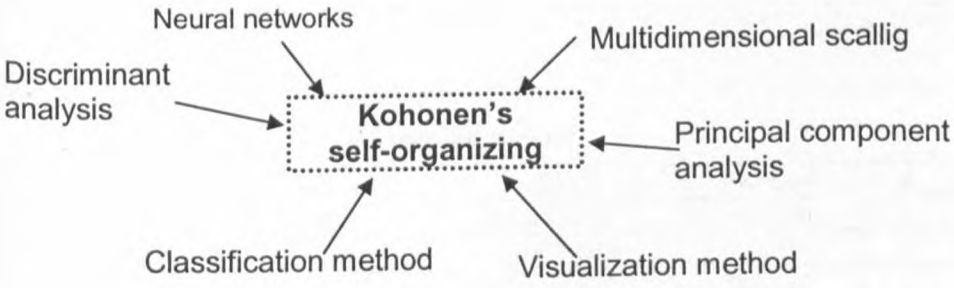

Figure 2. Self-organizing networks and other multivariate statistical analysis methods.
Source: Own research based on Kohonen [1997], Bock [2004].

SOMs can be treated as a branch of neural networks due to fact that data are processed sequentially, no calculations on data treated as numerical matrix is made. It is also a clustering method because mini-clusters are displayed in nodes of lattice and there is a classification process in background. It's quite obvious that this is a visualization method and due to fact that self-organizing maps reduce number of dimension there is an affinity to multidimensional scaling and principal components analysis. At last but not least supervised Kohonen's maps can be also used as discriminant method. Example of such use for symbolic data is presented in chapter IV.

III. SYMBOLIC VARIABLES AND SYMBOLIC OBJECTS

Symbolic data, unlike classical data, are more complex than tables of numeric values. While Table 1 presents usual data representation with objects in rows and variables (attributes) in columns with a number in each cell, table 2 presents symbolic objects with intervals, set and text data.

Table 1

Classical data situation

X	Variable 1	Variable 2	Variable3	...
1	1	108	11,98	
2	1,3	123	–23,37	
3	0,9	99	14,35	
...	...	...	...	...

Source: own research.

Table 2

Symbolic data table

X	Variable 1	Variable 2	Variable 3	Variable 4
1	(0,9;0,9)	{106;108;110}	11;98	{Blue;green}
2	(1;2)	{123;124;125}	–23;37	{light-grey}
3	(0,9;1,3)	{100;102;99;97}	14;35	{pale}
...	...	...	...	...

Source: own research.

Bock and Diday [2000] define five types of symbolic variables: single quantitative value, categorical value, interval, multivalued variable, multivalued variable with weights.

Variables in a symbolic object can also be, regardless of its type (Diday [2002]): taxonomic – representing hierarchical structure, hierarchically dependent, logically dependent.

There are four main types of dissimilarity measures for symbolic objects (Malerba *et al.* [2000], Ichino and Yaguchi. [1994]):

- Gowda, Krishna and Diday – mutual neighbourhood value, with no taxonomic variables implemented;
- Ichino and Yaguchi – dissimilarity measure based on operators of Cartesian join and Cartesian meet, which extend operators $\cup$ (sum of sets) and $\cap$ (product of sets) onto all data types represented in symbolic object,

- De Carvalho measures – extension of Ichino and Yaguchi measure based on a comparison function (CF), aggregation function (AF) and description potential of an object.
- Hausdorff distance (for symbolic objects containing intervals).

For symbolic data containing only interval-type variables Hausdorff distance and vertex-type distance (sum of squares of all distances between adequate vertices of n-dimensional hyper-cubes defined by n interval variables) is often used.

IV. CREATION OF SOMS FOR SYMBOLIC DATA IN FORM OF INTERVALS

El Golli, Conan-Guez and Rossi[2004] proposed an extension of original Kohonen's algorithm which allows creation of self-organizing maps for symbolic data containing intervals.

Two main innovations for SOMs for symbolic data are proposed. In original Kohonen algorithm in step 2 squared euclidean distance is used for assignment of actual object to the closest cluster. For data in form of intervals Hausdorff distance or vertex-type distance can be used.

Second change is that cluster prototypes are not points but hyper-cubes. Thus prototypes adjustment step (step 3) is repeated for each vertex of hyper-cube defined by intervals and formula (1) recalculates coordinates of each vertex separately.

V. EXAMPLES OF USE OF KOHONEN'S MAPS FOR SYMBOLIC OBJECTS

Symbolic data set containing information about car models has been used as input data for constructing self-organizing map. The result of this process shows figure 3.

The mini-cluster assignment is the following.

Cluster(1x1): Alfa 145, Vectra, Skoda Octavia, Cluster(1x2): Alfa 156, Rover 75; Cluster(1x3): Alfa 166, Lancia K, Mercedes Class C; Cluster(1x4): Aston Martin; Cluster(2x1): Audi A3; Cluster(2x2) : Audi A6; Cluster(2x3): Audi A8, Maserati GT; Cluster(2x4): Bmw serie 3; Cluster (3x1): Bmw serie 5; Cluster(3x2): Bmw serie 7, Mercedes SL, Mercedes Class E, Porsche; Cluster (3x3): Ferrari, Mercedes Class S; Cluster(3x4): Punto, Lancia Y; Cluster(4x1): Fiesta, Nissan Micra, Corsa, Twingo, Rover 25, Skoda Fabia; Cluster(4x2): Focus, Passat; Cluster(4x3): Honda NSK; Cluster(4x4): Lamborghini.

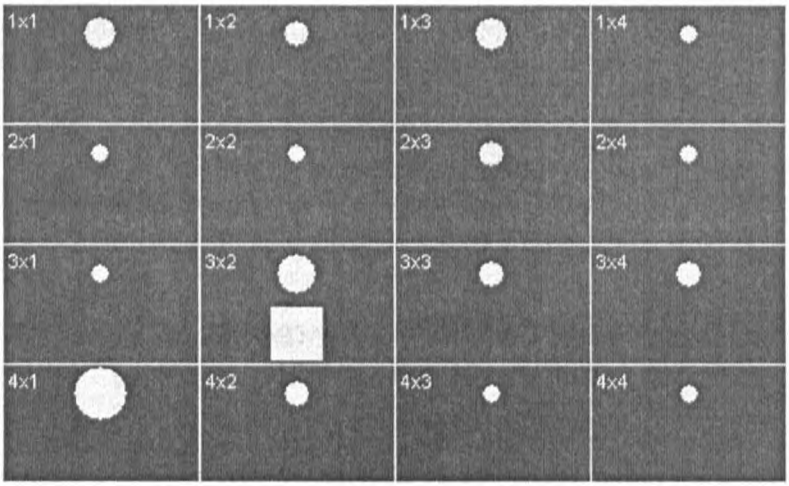

Figure 3. Self-organizing map for car.sds data set.
Source: own calculation with SODA 2.5 software, file car.sds comes from symbolic data repository http://www.ceremade.dauphine.fr/~touati/sodas-pagegarde.htm.

In second example SOMs has been used for discriminant analysis. From 177 objects sets containing information about french wines (taken from http://www.ceremade.dauphine.fr/~touati/sodas-pagegarde.htm) 120 objects has been treated as training set and 57 as test set. The map after learning process is showed on figure 4.

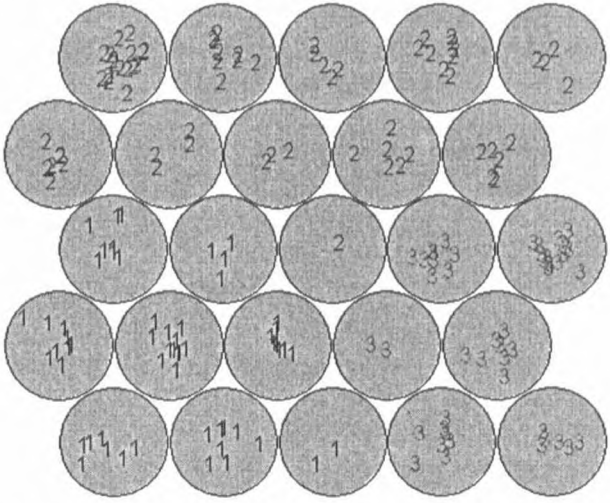

Figure 4. Self-organizing map for wine.sds data set after learning process.
Source: own calculations.

Table 3 is contingency table between actual and predicted class assignment.

Table 3

Contingency table for wine.sds file

	1	2	3
1	18	0	0
2	1	21	2
3	0	0	15

Source: Own calculations.

Error ratio for this prediction is 5,2% and Rand and corrected Rand of class agreement between real and predicted cluster structures is 0,92 and 0,84.

VI. FINAL REMARKS

Kohonen self-organizing maps can be, after small modifications of original algorithm, adapted for symbolic data in form of intervals. It is visualization method for this kind of data as well a clustering method. Supervised SOMs can also be treated as a discriminant analysis technique for symbolic interval data.

Still an open issue, worth further development is how to adapt Kohonen algorithm for other symbolic data type (nominal and multi-nominal, categorical data, distributions).

REFERENCES

Bock H.-H., Diday E (Eds.) (2000), *Analysis of symbolic data. Explanatory methods for extracting statistical information from complex data*, Springer Verlag, Berlin.

Bock H.-H. (2003), Clustering algorithms and Kohonen maps for symbolic data. *Journal of the Japanese Society of Computational Statistics,* 15.2, 217–229.

Diday E. (2002), An introduction to symbolic data analysis and the SODAS software, Journal of Symbolic Data Analysis, Vol. 1.

El Golli A., Conan-Guez B., Rossi F. (2004), Self Organizing Map and Symbolic Data. *Journal of Symbolic Data Analysis*, Vol. 2.

Ichino M., Yaguchi H. (1994), *Generalized Minkowski Metrics for Mixed Feature–Type Data Analysis*, „IEEE Transactions on Systems, Man, and Cybernetics", Vol. 24, No. 4, 698–707.

Kohonen T. (1997), *Self-organizing maps*, Springer-Verlag, Berlin.

Malerba D., Espozito F, Giovalle V., Tamma V. (2001), Comparing Dissimilarity Measures for Symbolic Data Analysis, *New Techniques and Technologies for Statistcs* (ETK–NTTS'01), 473–481.

Verde R.(2004), Clustering Methods in Symbolic Data Analysis, *Classification, Clustering and Data Mining*, Berlin-Springer-Verlag, 299–318.

Andrzej Dudek

SAMOORGANIZUJĄCE SIĘ MAPY KOHONENA DLA OBIEKTÓW SYMBOLICZNYCH

Wizualizacja danych w postaci diagramów i poszukiwanie w tych diagramach struktur, klas, trendów, zależności itp. jest jednym z głównych zadań wielowymiarowej analizy statystycznej. W przypadku danych symbolicznych (to jest danych reprezentowanych w postaci liczb, przedziałów liczbowych, zbiorów kategorii, czy zbiorów kategorii z wagami) wersje znanych metod takich jak analiza czynnikowa, czy analiza składowych głównych mogą być stosowane po pewnych modyfikacjach.

Alternatywną metodą wizualizacji danych są samoorganizujące się mapy Kohonena. Zamiast wyświetlać $k = 1, ..., n$ obiektów w dwuwymiarowej przestrzeni w postaci punktów czy prostokątów obiekty są najpierw dzielone na m mini-klas a następnie te mini-klasy przyporządkowywane są wierzchołkom prostokątnej kraty na płaszczyźnie w taki sposób aby „podobne" mini-klasy były przyporządkowane sąsiednim wierzchołkom kraty.

W artykule przedstawiony został algorytm tworzenia map Kohonena dla danych symbolicznych oraz przykłady jego zastosowania dla danych symbolicznych pochodzących z repozytorium http://www.ceremade.dauphine.fr/~touati/sodas-pagegarde.htm.