New advances in Symbolic Data Analysis and Spatial Classification

Edwin DIDAY

Abstract:
The usual Data mining model is based on two parts: the first concerns the units (called here "individuals"), the second, contains their description by several standard variables. The Symbolic Data Analysis model needs two more parts: the first concerns units called "concepts" and the second concerns their "description". Each concept is associated to a category of any categorical variable of the given data (as for example a class variable). The concepts are characterized by a set of properties using the initial variables called "intent" and by an "extent" defined by the set of individuals which satisfy these properties. These concepts are described by "symbolic data" which are standard categorical or numerical data and moreover interval, histograms, sequences of values, etc. These new kind of data allows keeping the internal variation of the extent of each concept. Then, new knowledge can be extracted from this model by new tools of Data Mining extended to concepts considered as new units. Among these tools, Spatial Classification allows a graphical visualization of the given concepts on a grid and at different level of generalization organized by a spatial hierarchy or pyramid (allowing overlapping clusters). The SYR software for Symbolic Data Analysis is a professional software which has been developed by SYROKKO company after the academic SODAS software developed by two EUROPEAN projects until 2003.
Now we summarize the pyramidal classification model and theory.

Keywords:
Symbolic Data Analysis, Data Mining, Spatial Classification, Pyramidal clustering,

SPATIAL CLASSIFICATION GRAPHICAL MODEL

1. INTRODUCTION

The aim of a spatial classification is to position the units on a spatial network and to give simultaneously a set of structured classes of these units "compatible" with the network. We introduce the basic needed definitions: compatibility between a classification structure and a tessellation, (m, k)-networks as a case of tessellation, convex, maximal and connected subsets in such networks, spatial pyramids and spatial hierarchies. Like Robinsonian dissimilarities induced by indexed pyramids generalize ultrametrics induced by indexed hierarchies we show that a new kind of dissimilarities called "Yadidean" induced by Spatial Pyramids generalize Robinsonian dissimilarities. We focus on spatial pyramids where each class is a convex for a grid, and we show that there are several one-to-one correspondences with different kinds of Yadidean dissimilarities. These new results produce also, as a special case, several one to one correspondences between spatial hierarchies (resp. standard indexed pyramids) and Yadidean ultrametrics (resp. Robinsonian) dissimilarities. Qualities of spatial pyramids and their supremum under a given dissimilarity are considered. We give a constructive algorithm for convex spatial pyramids illustrated by an example. We show finally on a simple example that Spatial pyramids on symbolic data can produce a

geometrical representation of conceptual lattices of "symbolic objects".

2. MAIN THEOREMS

Indexed hierarchies and ultrametrics yield a one-to-one correspondence shown by Johnson[8] and Benzecri [2]. Diday [6] has shown a one-to-one correspondence between indexed clustering pyramids and Robinsonian dissimilarities which generalize the one-to-one correspondence between indexed hierarchies and ultrametrics. These one to one correspondences have been studied by several authors, for example Bertrand, Janowitz [3], Bertrand [4]. In order to build a clustering pyramid, several algorithms have been proposed by Diday [6], Aude [1] for the standard case of classical variables and by Brito [5], Rodriguez [9] for the symbolic data case.

We introduce a case of tessellation called (m, k)-network. It is a grid when m = k = 4. When the tessellation is reduced to a chain with edges of equal size on a straight line we say that it is a (2, 2)-network. Spatial pyramids are based on a graph defined by a m/k-network for which each cluster of the pyramid is "convex", "maximal" or "connected". The "compatibility" between an order O and a dissimilarity which is expressed by a Robinsonian matrix ordered by O, is generalized to the "compatibility" between a dissimilarity and a grid M expressed by a "Yadidean matrix" "ordered" by M. "Yadidean dissimilarities" generalize Robinsonian dissimilarities as a Yadidean dissimilarity is a Robinsonian dissimilarity in the case of a (2, 2)-network. The one-to-one correspondence given in Diday [7] between a family of " indexed spatial pyramids" and a **family** of Yadidean dissimilarities is generalized to one-to-one correspondences between several kinds of equivalence classes of indexed spatial pyramids and several kinds of Yadidean dissimilarities called "large", "strict", "weakly large", "weakly strict". We extend standard hierarchies to spatial pyramids and ultrametrics to Yadidean ultrametrics. Then, we show that these results lead to several kinds of one to one correspondences between indexed hierarchies and ultrametrics, between indexed pyramids and Robinsonian dissimilarities and between spatial hierarchies and Yadidean ultrametrics. We show that the supremum of the set of Yadidean ultrametrics lower than a given dissimilarity is a Yadidean dissimilarity. We give a constructive algorithm for convex spatial pyramids illustrated by an example. Finally, we show by a simple example that spatial pyramids can give a geometrical representation of a conceptual lattice. In figure 1 we give an example of spatial pyramid and a tool for a graphical interpretation of each level of the pyramid. This work has been done by the project SEVEN sponsored by the ANR “Agence Nationale pour la Recherche” and directed by EDF (Clamart) from 2005 to 2008, with the participation of the LIMSI, IINRIA, and CEREMADE. For the CEREMADE, M. Touati and M. Rahal have actively participated.

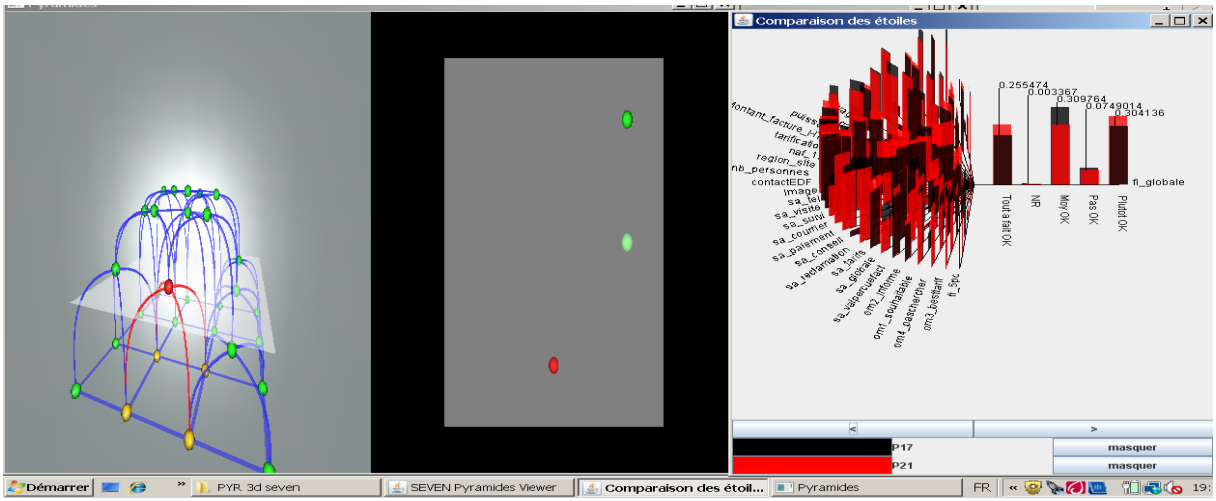

Figure 1 Graphical visualization of a spatial classification by a 3-D pyramid

3. CONCLUSION

A wide field of research is opened by extending the results already obtained in standard hierarchies and pyramids to spatial pyramids compatible with a grid, then, by extending these new results to other kinds of classes (for instance, of maximal or connected classes instead of convex), to other kinds of grids (as triangular or hexagonal) and to multidimensional grids. For instance, in the case of a cubic grid we can obtain a 3-D Yadidean dissimilarity defined by blocks which are 2-D Yadidean dissimilarities increasing from the main diagonal in rows and columns. In that way, we can go more generally, from a n-D Yadidean dissimilarity to a (n+1)-D one. In the 3-D Yadidean dissimilarity case, the classes of the associated classification structure are volumes as they merge cells of the 3-D grid. They form a partitioning or an overlapping of the 3-D grid depending on the fact that the 3-D associated Yadidean dissimilarity is "ultrametric" or not, etc. Many other directions remain open, such as how to get the closest Yadidean dissimilarity of a given dissimilarity and what is the statistical distribution of a quality criterion between a given dissimilarity and different kinds of Yadidean dissimilarity (weakly large, large, weakly strict, strict…) ?. It is possible to do a spatial classification of spatial units, for example, what is the three dimensional spatial pyramidal structure of the concepts of our brain by using as input a dissimilarity between their dictionary definition. What is the three dimensional spatial pyramidal structure of the stars of the sky by using their distances as input?

References:

Recent Books

L. Billard, E. Diday (2006) "Symbolic Data Analysis: conceptual statistics and data Mining". Wiley. ISBN 0-470-09016-2. 351 pages.

E. Diday, M. Noirhomme (editors and co-authors) (2008) "Symbolic Data Analysis and the SODAS software" 457 Pages. Wiley. ISBN 978-0-470-01883-5.

P. Brito, P. Bertrand, G. Cucumel, F. De Carvalho (editors) (2007). Selected contributions in Data Analysis and Classification. ISBN 978 3 540 73558 8 Springer Berlin Heidelberg New York.

Advised recent paper on the theory

E. Diday (2008). Spatial classification. DAM (Discrete Applied Mathematics) Volume 156, Issue 8, Pages 1271.

Other books and papers:

[1] H.H. Bock, E. Diday (eds.): Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data . Springer Verlag, Heidelberg, 425 pages, ISBN 3-540-66619-2.

[2] J.P. Benzecri, L'Analyse des données: la Taxinomie, Vol. 1, Dunod, Paris, 1973.

[3] P. Bertrand, M.F. Janowitz, Pyramids and weak hierarchies in the ordinal model for clustering. Discrete Applied Mathematics. 122, pp. 55-81, 2002.

[4] P. Bertrand Structural properties of pyramidal clustering. Dimacs Ser. Theor Comput. Sci. 19 35-53, 1995.

[5] P. Brito Order structure of symbolic assertion objects. IEEE TR. on Knowledge and Data Engineering Vol.6, n° 5, October, 1994.

[6] E. Diday, Orders and Overlapping clusters in pyramids. In J. De Leeuw, et al., (Eds.). Multidimensional Data Analysis, DSWO Press , Leiden, pp. 201-234. 1986.

[7] E. Diday, " Spatial Pyramidal Clustering Based on a Tessellation". Proceedings IFCS'2004. Proceedings of the Meeting of the International Federation of Classification Societies. Illinois Institute of Technology, Chicago, 15-18 July 2004, D. Bank and al. Editor. Springer Verlag, pp. 105-120. 2004.

[8] E. Diday (2008) Spatial classification. DAM (Discrete Applied Mathematics) Volume 156, Issue 8, Pages 1271-1294.

[9] S.C. Johnson, Hierarchical clustering schemes, Psychometrika 32 pp. 241-254, 1967.

[10] K.Pak, M.C.Rahal et E.Diday. Elagage et aide à l'interprétation symbolique et graphique d'une pyramide. Congrés d'extraction et gestion des connaissances, EGC 18-21 Janvier 2005 Paris, Editions Cepadues