Guest Editorial

Symbolic and spatial data analysis: Mining complex data structures

Paula Brito[a] and Monique Noirhomme-Fraiture[b]
[a]*Faculdade de Economia, University of Porto, Rua Dr. Roberto Frias, 4200-464 Porto, Portugal*
E-mail: mpbrito@fep.up.pt
[b]*Institut d'Informatique, Facultés Universitaires Notre Dame de la Paix, Rue Grandgagnage, 21,*
B-5000 Namur, Belgium
E-mail: monique.noirhomme@info.fundp.ac.be

Nowadays, researchers from different areas must face the growing complexity of the information available, which is characterised by the generally high dimensions of databases and their structure, resulting from the observation of multivariate phenomena in different state/space occasions. As a consequence, there is an increasing interest in extracting knowledge from large collections of data. Data mining, knowledge discovery in data bases, intelligent data analysis are some of the terms adopted to identify parallel streams of work aiming to support humans in extracting previously unknown, valid, potentially useful and understandable patterns in the data. Most studies in these areas have until recently focused on a relatively simple representation of data: a database relation, or a standard data table, or a set of points in a feature space. In fact, the relational model is clean and simple, and a relational table can be easily mapped into the mathematical concept of matrix. Moreover, many data analysis applications concern administrative data, which are easily represented by this model. With the advent of the "information age", we have witnessed to a dramatic growth of applications in government, business and education, many of which are sources of various data, organised in different structures and formats. The chances that computers have provided have enlarged the meaning of "data", have defined new sorts of problems in knowledge discovery, and have led to the development of completely new classes of models and data analysis algorithms. Object oriented databases, and, more recently, object-relational databases allow for the manipulation of data with complex structures, which then require novel methodologies of analysis.

1. Symbolic data analysis

There is an increasing need to extend standard exploratory, statistical and graphical data analysis methods to the case of more complex data, that go beyond the classical framework. This is the case of data concerning more or less homogeneous classes or groups of individuals (second-order objects or macro-data), instead of single individuals (first-order objects or micro-data). The extension of classical

1088-467X/06/$17.00 © 2006 – IOS Press and the authors. All rights reserved

data analysis techniques to the analysis of second-order objects is one of the main goals of a novel research field named "symbolic data analysis". Symbolic Data Analysis allows defining concepts by a query on a database, aggregate initial data in order to describe these concepts (as symbolic data) and then apply analysis methods to extract knowledge from the set of modelled concepts. Symbolic data extend the classical tabular model, allowing multiple, possibly weighted, values for each descriptive attribute which allow representing variability and/or uncertainty present in the data. Symbolic Data Analysis methods include univariate descriptive methods, clustering, decision-tree, discrimination, regression and factorial analysis techniques, which allow analysing symbolic data tables. Symbolic data occur in many situations, for instance in summarising huge sets of data or in describing the underlying concepts (a town, a socio-demographic group, a scenario of accidents) of a database. It also finds an important application field in official statistics; since by law, NSI's are prohibited from releasing individual responses to any other government agency or to any individual or business, data are aggregated for reasons of privacy before being distributed to external agencies and institutes. Symbolic Data Analysis provides useful tools to analyse such aggregated data. Symbolic Data Analysis underwent great improvement with the European projects "Symbolic Official Data Analysis System (SODAS)" and "Analysis System for Symbolic Official Data (ASSO)"; as the result of these projects a software package SODAS has been developed.

2. Spatial data analysis

With the exponentially growing use of geographic information systems (GIS) to store, manipulate and visualize geo-coded information, it is increasingly important to understand the particular nature of geographic data and the specialized techniques required for its analysis. There is increasing interest in studying how techniques for the analysis of spatial data can be effectively applied in a GIS environment, such as the study of spatial patterns and spatial autocorrelation, detection of clusters, outliers and any other relationships that pertain to the absolute and relative location of observations. Common applications of spatial data analysis techniques in the social sciences range from the discovery of crime clusters, hot spots and the detection of disease clusters, to spatial autocorrelation of demographic variables and regression models for real estate analysis. Other applications concern public health services searching for explanations of disease clusters, environmental agencies assessing the impact of changing land use patterns on climate change, geo-marketing companies doing customer segmentation based on spatial location, etc. For supporting this type of analysis, most contemporary GIS have only very basic spatial analysis functionalities; many are confined to analysis that involves descriptive statistical displays, such as histograms or pie charts. Data mining, which is the partially automated search for hidden patterns in large databases, offers great potential benefits for the applied GIS based decision making that takes place in public and private sector organizations.

3. Content

This special issue includes five papers which constitute updated and extended versions of papers selected from those presented at the Workshop on "Symbolic and Spatial Data Analysis: Mining Complex Data Structures" chaired by the guest editors of this issue, in Pisa in September 2004. The workshop was organized within the framework of the 15th European Conference on Machine Learning (ECML'04) and the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'04).

The main goal of this workshop was to bring together researchers from different communities such as machine learning, data analysis, symbolic data analysis and data mining to promote discussion and the development of new ideas and methods to deal with such complex structured data.

To give the reader a better insight of this issue, we summarize the main ideas behind the individual papers in the following.

The first paper, “Classification of Symbolic Objects: A Lazy Learning Approach”, by Annalisa Appice, Claudia D’Amato, Floriana Esposito and Donato Malerba, presents a lazy-learning approach that extends a traditional distance weighted k-Nearest Neighbor classification algorithm to symbolic data. The proposed method has been implemented in the system SO-NN (Symbolic Objects Nearest Neighbor) and evaluated on symbolic datasets.

The second paper, “The criterion of Kolmogorov-Smirnov for binary decision tree: Application on interval valued variables” from C. Mballo and E. Diday, proposes to extend classification by binary tree to the case of interval valued data, and this extension is achieved by applying Kolmogorov-Smirnov criterion to interval valued variables.

The third paper, “An Algebraic Method for Compressing Symbolic Data Tables”, by Yannis Tzitzikas, proposes a novel technique for compressing a symbolic data table using the recently emerged Compound Term Composition Algebra. One advantage of CTCA is that the closed world hypotheses of its operations can lead to a remarkably high “compression ratio”. The compacted form apart from having much lower storage space requirements, allows designing more efficient algorithms for symbolic data analysis.

The fourth paper, “Materialized aR-Tree in Distributed Spatial Data Warehouse”, by Marcin Gorawski and Rafal Malczok, presents a Spatial Data Warehouse system that is used for aggregation and analysis of huge amounts of spatial data. The data is generated by utilities meters communicating via radio. In order to provide sufficient efficiency for the system, authors propose data and workload distribution as well as advanced indexing techniques. The system is based on a cascaded star model, which is a spatial development of a standard star schema and contains interconnected and often nested star schemas. Thanks to an available memory evaluating mechanism the system is very flexible in the field of aggregates accuracy. They also have implemeted indexing structure updating mechanism. Basing on the wide variety of test results, they prove that a distributed system significantly surpasses the centralized version in terms of efficiency.

The fifth paper, “Towards symbolic mining of images with association rules: Preliminary results on textures”, by Matjaz Bevk and Igor Kononenko, presents new textural features which are based on association rules. They give a texture representation, which is an appropriate formalism, that allows straightforward application of association rules algorithms. This representation has several good properties like invariance to global lightness and invariance to rotation. Association rules capture structural and statistical information and are very convenient to identify the structures that occur most frequently and have the most discriminative power. The results from their experiments show that this representation gives comparable results to standard texture descriptions and better results than general image descriptions.

Although the papers in this special issue do not cover the wide range of topics of the analysis of complex data, they certainly give the reader an idea of some of the challenging problems arising in this field. This is a most promising research area, where important developments are to be expected. We hope to have contributed to stimulate added interest and further research.

Acknowledgements

We gratefully acknowledge the remarkable work of the members of the program committee of the workshop. They helped us in selecting the contributions for the workshop and in reviewing the papers for this special issue and thereby improve their quality. They are: Annalisa Appice, Patrice Bertrand, Lynne Billard, Gilles Bisson, Hans-Hermann Bock, Carmen Bravo, Marie Chavent, Francisco De Carvalho, Edwin Diday, K. Chidananda Gowda, Georges Hébrail, Manabu Ichino, Alípio Jorge, Yves Lechevallier, Petko Valtchev, Rosanna Verde.