ECML/PKDD'02 Workshop on "Mining Official Data" (MOD'02)

19th or 20th August 2002
Helsinki, Finland

before the
13th European Conference on Machine Learning (ECML'02)
6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02)
Helsinki, Finland

under the auspices of the
KDnet - the European Network of Excellence in Knowledge Discovery

Technical description

In statistics, the term "*official data*" denotes data collected in censuses and statistical surveys by National Statistics Institutes (NSIs), as well as administrative and registration records collected by government departments and local authorities. They are used to produce "official statistics" for the purpose of making policy decisions, and to facilitate the appreciation of economic, social, demographic, and other matters of interest to the governments, government departments, local authorities, businesses, and to the general public. For instance, population and economic census information is of great value in planning public services (education, fund allocation, public transport), as well as in private businesses (placing new factories, shopping malls, or banks, as well as marketing particular products). Moreover, survey data on specific topics, such as labour force, time use, household budget, are regularly collected by NSIs to keep updated information on some economic and social phenomena.
The application of data mining techniques to official data has great potential in supporting good public policy and in underpinning the effective functioning of a democratic society. Nevertheless, it is not straightforward and requires a challenging methodological research, which is still in an initial stage. In particular, to develop successful applications of data mining techniques to official data, the following issues must be dealt with:

a. ***Aggregated data***. NSIs make a great effort in collecting census data, but they are not the only organizations that analyse them: data analysis is often done by different institutes. By law, NSIs are prohibited from releasing individual responses to any other government agency or to any individual or business, so data are aggregated for reasons of privacy before being distributed to external agencies and institutes. Data analysts are confronted with the problem of processing data that go beyond the classical framework, as in 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 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".
b. ***Data quality***. There are different ways to determine data quality, including the percentage of errors in data and the use of an unbiased sampling procedure. One problem with official data is the human error involved in inputting. Some outlier detection techniques developed in data mining can be used to find errors in collected data. Data can also be missing, which is the

biggest problem with official data. In this case, clustering methods can be used to replace missing values. Finally, data mining techniques can also be used to control the sample bias, before disseminating official aggregated data used in further processing.

c. ***Timeliness***. This can be considered another aspect of data quality. Public and private institutions are currently urged to reduce the delay between the time of data collection and the moment in which decisions are made according to some statistical indicators. A typical example is the inflation rate computed by the European Institute of Statistics (Eurostat) and the decision made by the Central Bank of Europe (BCE) on the tax rate. A timely delivery of data analysis results may involve the synthesis of new indicators from official data, the design of different infrastructures for timely data collection, or the application of 'anytime algorithms', which provide the data miner with a ready-to-use model at any time after the first few examples are seen and guarantee a smooth quality, increasing with time.
d. ***Geo-referentiation***. The practice of geo-referencing census data has increasingly spread over the last few decades and the techniques for attaching socio-economic data to specific locations have markedly improved at the same time. In the UK, for instance, household expenditure data are provided for each enumeration district (ED), the smallest areal unit for which census data are published. At the same time, vectorized boundaries of the 1991 census EDs enable the investigation of socio-economic phenomena in association with the geographical location of EDs. These advances cause a growing demand for more powerful data analysis techniques that can link population data to their spatial distribution.
e. ***Metadata***. In statistics, metadata concerns the information used for the most correct understanding of statistical data and their related analysis. They mainly refer to explanations-definitions-procedures that are followed from the designing phase up to the phase of a publication of survey's results. Examples of metadata are the various statistical populations, sampling techniques, definitions of nomenclatures, classifications, monetary units and so on. The basic use of metadata is in interpreting and validating data, as well as in finding and accessing relevant information. However, metadata can also be used for analysis purposes. This is notably the case of research studies performed on government and official statistics, since macro data are of little use to any data consumer if they are not accompanied by additional information, such as what they represent, how they were collected and manipulated and so on.

Topics

The workshop will maintain a balance between theoretical issues and descriptions of case studies to promote synergy between theory and practice. *Research contributions are welcome even though they have not been tested on "official data" but have a clear relation with some of the research issues reported in the technical description.* Topics of interest include, but are not limited to:

- Methodologies and policies for the analysis of official data
- Confidentiality protection in data mining
- Mining aggregated data
- Symbolic data analysis
- Data mining for quality control in data capture and transformation
- Data mining techniques for outlier detection
- Data mining techniques for qualitative comparison of statistics
- Infrastructures for timely collection/delivery of official data/statistics
- Anytime data mining algorithms for timely delivery of official statistics

- Spatial data mining of official/business data
- Infrastructures for the provision of metadata
- Use of meta-data in data mining techniques
- Application of meta-data to the validation of data mining results
- Case studies of mining official data
- Descriptions of official data sources and related data mining problems.

Workshop structure and attendance

The workshop aims to be a highly communicative meeting place for researchers working on similar topics, but coming from different communities. In order to achieve these goals, the workshop will consist of two invited talks, followed by short presentations and longer discussions. Each author will be encouraged to read another accepted paper and to comment on it after the original talk has been given.

All ECML/PKDD'02 MOD workshop participants must also register for the main ECML/PKDD conference. Workshop attendance will be limited to registered participants.

Submission Procedure

Authors are invited to submit original research contributions or experience reports in English. Submitted papers must be unpublished and substantially different from papers under review. Papers that have been or will be presented at small workshops/symposia whose proceedings are available only to the attendees may be submitted.
Papers should be double-spaced and no longer than 5000 words (about 12 single-spaced pages). Papers should be sent electronically (postscript or pdf) not later than **May 24, 2002** to

mod@di.uniba.it

Papers will be selected on the basis of review of full paper contributions. Authors should make certain that the data mining techniques they describe deal with the special issues that are associated with official data. Notification of acceptance will be given by June 14, 2002. Final camera-ready copies of accepted papers will be due by **June 28, 2002**. The proceedings will be printed by the ECML/PKDD organizers and distributed at the workshop. A web-publication of the proceedings is expected after the conference.

Style Guide

There is a joint paper style for the proceedings of all ECML/PKDD workshops. Submitted papers should be formatted according to the Springer-Verlag Lecture Notes in Artificial Intelligence guidelines. Authors' instructions and style files can be downloaded from http://www.springer.de/comp/lncs/authors.html.

Important Dates

Submission deadline:	May 24, 2002
Notification of acceptance:	June 14, 2002
Camera ready copies of papers:	June 28, 2002
Workshop:	August 19/20, 2002

Organizing Committee

This workshop will be organized by:

Paula Brito, Faculty of Economics, University of Porto, Portugal
Donato Malerba, Department of Informatics, University of Bari, Italy

Program Committee

Timo Alanko, Statistical R&D Unit, Statistics Finland, Helsinki, Finland
Edwin Diday, CEREMADE, Paris-9 Dauphine University, Paris, France
Floriana Esposito, Department of Informatics, University of Bari, Italy
Paulo Gomes, National Institute of Statistics (INE), Lisbon, Portugal
Haralambos Papageorgiou, Department of Mathematics, University of Athens, Athens, Greece
Willi Klösgen, Fraunhofer Institute for Autonomous Intelligent Systems, Sankt Augustin, Germany
Carlos Marcelo, National Institute of Statistics (INE), Lisbon, Portugal
Michael May, Fraunhofer Institute for Autonomous Intelligent Systems, Sankt Augustin, Germany
Monique Noirhomme, Institut d'Informatique, University Notre-Dame de la Paix, Namur, Belgium
Mireille Summa, CEREMADE, Paris-9 Dauphine University, Paris, France
Ian Turton, Centre for Computational Geography, University of Leeds, Leeds, UK,

Related events

KDD-2002
The 2002 IAOS Conference on Official Statistics and the New Economy