ECML - 2000 WORKSHOP

DEALING WITH STRUCTURED DATA
IN MACHINE LEARNING AND STATISTICS

Final report

Organizing Committee:

Paula Brito Univ. Porto, Portugal
Joaquim Costa Univ. Porto, Portugal
Donato Malerba Univ. Bari, Italy

Place and date: Barcelona, May 30, 2000

1. Introduction

Nowadays, 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 focussed 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.
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 data, henceforth denoted as "structured data." Therefore, the workshop presented the great potentialities and difficulties of all multidisciplinary events that try to put in touch people with different background, experience and terminology.

2. Where is the structure in structured data?

The different presentation at the workshop made clear that data may present some structure at different levels.
A first type of structured data is represented by taxonomic attributes, that is attributes whose categories are ordered in a rooted hierarchical tree, called taxonomy. In this case the structure is *in the attribute domain*.

In data analysis taxonomic data led to the definition of suitable similarity indices to be used in categorisation problems. In data mining taxonomies are used to support generalisation-based knowledge discovery or attribute-oriented induction in order to reduce the computational complexity of the mining algorithms, while in machine learning taxonomies simply define some form of background knowledge to be used during the learning process.
A different, but someway related, form of structure in the attribute domain is that of *relational variables/attributes*, as they are referred to in the field of data analysis. In this case, a dissimilarity matrix is defined on the value sets. This dissimilarity matrix is used to define the dissimilarity/similarity between objects described by one or more relational attribute. It is noteworthy that the term "relational" used in data analysis has a quite different meaning from that attributed by people working in machine learning.

Dependencies may also exist *between variables*. They define another form of "structure" in the data that goes a step further than the standard data table. Variable dependencies may be logical (e.g. if colour is blue then type is river), causal (e.g. if driving speed is high, then nb. of accidents is high with probability 0.8) or the so-called "mother-daughter" relations, expressing that the applicability of one variable depends on the values taken by another one e.g. (if gender is male then nb .of pregnancies is non-applicable).

The invited talk by Prof. Lerman, who has been working in the area of data analysis for many years, focussed on these classes of "structured data."

Another type of structured data is represented by *aggregated data* used to represent classes or groups of individuals. These aggregated data, which are described by set-valued variables and modal variables, are called *symbolic data* and the extension of standard statistical methodology to analyse such data is the main goal of a recently developed area called symbolic data analysis. For different reasons, data warehouses and census data available at national offices of statistics, are two great potential sources of aggregated data. In symbolic data analysis, logical and mother-daughter dependencies are also investigated, thus introducing an additional degree of structure on the data.

The talk by Brito and Malerba showed that aggregated data represent a new promising research direction for machine learning, which has already developed quite sophisticated tools for controlled generalisation. Moreover in the talk by Tamma several metrics on aggregated data have been presented and compared: they are the basic tools on which new data analysis techniques can be defined.

Both taxonomic and symbolic data are extension of classical data tables, where attributes are either taxonomic or multi-valued or multi-modal, possibly with dependencies. By representing observations (or objects) as rows of the data table and attributes as columns, we can easily see that all types of "structures" presented above affect either a single column, or multiple columns, but they never express some kind of dependence between rows, that is objects.

A more complex representation is given by first-order logic, where both attributes of single objects and *relations between objects* are represented. Statistical data analysts have hardly tackled this sort of structured data, primarily because the independence

assumption of observations, which is fundamental in statistics, does no longer apply. Upgrading statistical data analysis tools, from dissimilarity measures to more complex classification methods such as Naive Bayesian, is still an open problem and a promising direction of research.

The invited talk by Prof. Esposito and the talk by Flach provided some insights on how such upgrading of statistical analysis tools can be reached either by integrating different techniques according to a truly multistrategy perspective, or by computing a (posterior) probability for logical formulae.

Object-oriented representations offer another example of complex "structure" with problems of cyclicity on concept dependencies. The talk by Valtchev illustrated some difficulties arisen in computing similarities/ dissimilarities on instances of classes of an object-oriented database.

Finally, the structure on data makes it possible to define abstractions between target relations that are to be induced from examples. Given an abstraction relation between two relations, the two relations have to satisfy some constraints in order to be consistent with respect to the given abstraction. In his invited talk, Prof. Bratko has explained how these consistency constraints can be used to guide the search among possible hypotheses at different levels of abstraction.

In the workshop, different domains in which structured data arise have also been reported, such as official statistics for the handling of census data and survey data, where questions are often dependent on each other, data warehouses, GIS applications, XML documents or genomic databases.

3. The workshop in figures

Three invited talks were given in the workshop:

- I.C. Lerman from the University Rennes I and IRISA, France, presented a talk on "Comparing taxonomic data" ;
- Bratko from the University of Ljubljana, Slovenia, talked about "Abstractions between learning problems based on abstractions between structured data" ;
- F. Esposito from the University of Bari, Italy, presented a talk entitled "Inductively learning from numeric and symbolic data: A multistrategy view".

Other contributions were:

- P. Brito and D. Malerba – "Symbolic data analysis and machine learning: bridging the gap" ;
- P.A. Flach – "Decomposing probability distributions on structured individuals"
- D. Malerba, L. Sanarico and V. Tamma – "A comparison of dissimilarity measures for Boolean symbolic data"
- P. Vatchev and R. Missaoui – "Exploration of complex objects structure for knowledge discovery"

The workshop was intended to be a genuinely interactive event and not a mini-conference. Thus, ample time was allotted for general discussion. We registered more than twenty participants and a true interest in the audience, which considered the topics of the workshop quite "hot", as also witnessed by two related events in North America

(ICML Workshop on " Attribute-Value and Relational Learning: Crossing the Boundaries", June 2000; AAAI Workshop on Learning Statistical Models from Relational Data, July 2000).

Acknowledgements

Thanks to ECML'2000 co-chairs and LIACC – Univ. Porto for supporting the organisation of this workshop and the European Network of Excellence in Machine Learning (MLNET) for the economical support.