Soft Comput (2007) 11:397–399
DOI 10.1007/s00500-006-0105-3

EDITORIAL

Special issue on soft computing for information mining

Ulrich Bodenhofer · Eyke Hüllermeier · Frank Klawonn · Rudolf Kruse

Published online: 20 June 2006

The rapid development of computers has not only led to an enormous increase in computing power but also to cheap mass storage media. As a result, collecting large amounts of data has nowadays become a routine in science, production, business, and commerce. However, to access the information contained in the raw data, appropriate data analysis techniques are needed. The traditional approach in statistics is to generate hypotheses first and to test them afterwards: One starts by defining the question to be answered, then designs an appropriate experiment and collects the data, and finally analyzes the data against the background of the original goals. This process is more or less turned upside down when permanently collecting huge amounts of data, mostly without having a concrete hypothesis in mind. As a consequence, there is a high demand for exploratory (instead of inferential) data analysis techniques that deliver a quick insight into the data characteristics and provide easily interpretable results.

As a response to the above development, along with the limited human capabilities in analyzing and exploiting large amounts of data, the field of knowledge discovery in databases (KDD) has recently emerged as a new research discipline, lying at the intersection of statistics, machine learning, data management, and related areas. According to a widely accepted definition, KDD refers to the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable structure in data. The central step within the overall KDD process is *data mining* or, linguistically perhaps more correct, *information mining*, the application of computational techniques to the task of finding patterns and models in data. Still, KDD also involves further important steps, notably data preparation, data cleaning, incorporation of prior knowledge, and interpretation of data mining results.

A multitude of efficient algorithmic methods for finding "interesting" patterns and relationships in large data sets have been devised in recent years. These methods exploit the capability of computers to search huge amounts of data in a fast and effective manner. More often than not, however, the data to be analyzed is imprecise and afflicted with uncertainty. In the case of heterogeneous data sources such as text and video, the data might moreover be ambiguous and partly conflicting. Last but not least, patterns and relationships of interest are usually vague and match with the data at best approximately. Thus, in order to make the information mining process more robust or, say, "human-like", methods for searching and learning are needed that are tolerant toward imprecision, uncertainty, and exceptions,

U. Bodenhofer
Institute of Bioinformatics, Johannes Kepler University,
Altenbergerstr. 69, 4040 Linz, Austria
e-mail: ulrich.bodenhofer@jku.at

E. Hüllermeier (✉)
Department of Computer Science (ITI),
Otto-von-Guericke-University of Magdeburg,
Universitätsplatz 2, 39106 Magdeburg, Germany
e-mail: eyke.huellermeier@iti.cs.uni-magdeburg.de

F. Klawonn
Department of Computer Science,
University of Applied Sciences BS/WF,
Salzdahlumer Str. 46/48, 38302 Wolfenbüttel, Germany
e-mail: f.klawonn@fh-wolfenbuettel.de

R. Kruse
Department of Computer Science (IWS),
Otto-von-Guericke-University of Magdeburg,
Universitätsplatz 2, 39106 Magdeburg, Germany
e-mail: kruse@iws.cs.uni-magdeburg.de

have approximate reasoning capabilities and are able to handle partial truth.

Here is where "soft computing" comes into play. One of the main concerns of "soft" computing techniques, a collection of methodologies whose cornerstones are fuzzy logic, neural networks, and evolutionary algorithms, is to complement classical, "hard" computing techniques with properties of the aforementioned kind. For example, the capability of fuzzy sets to interface quantitative patterns with qualitative knowledge structures expressed in terms of natural language can improve the comprehensibility of extracted patterns considerably, which is a point of major importance in data mining. Fuzzy information granulation further allows for trading off accuracy against efficiency and understandability of models. Among other things, fuzzy sets can also be useful in data reduction, in dealing with incomplete and heterogeneous data, in modeling prior knowledge, or in interactive data mining, where the mining process is under partial control of the analyst. These flexible modeling and knowledge representation capabilities of fuzzy sets are nicely complemented by efficient learning and adaptation techniques developed in the field of neural information processing, as well as robust search and optimization strategies as offered by evolutionary algorithms. In fact, soft computing solutions are especially effective if the strengths of the different techniques are combined into hybrid systems.

In the context of information mining, it can be hoped that soft computing – possibly in conjunction with approaches from other fields such as statistics, data analysis, and machine learning – contributes to methods that combine the complementary searching, reasoning and pattern recognition capabilities of both computers and humans in an optimal manner. This is the main motivation for devoting a special issue to the application of soft computing methods in information mining. The latter grew out of a couple of events organized by the editors in recent years, notably a special track on "Fuzzy Methods in Learning from Data" at the 13th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2004) held in Budapest in July 2004 and a workshop on "Soft Computing for Information Mining" held as part of the 27th German Conference on Artificial Intelligence (KI-2004) that took place in Ulm in September 2004. In fact, some of the contributions in this issue are revised and extended versions of papers that have been presented on one of these occasions. Nevertheless, the present special issue was preceded by a regular and open call for papers, and all submissions have passed through a rigorous reviewing process.

The papers in this special issue contribute to different areas of information mining. The first group of papers introduces new techniques for supervised learning from data. Alcalá, Alcalá-Fdez, Gacto and Herrera propose a new type of fuzzy rules that incorporate small shifts as well as narrowing and broadening of the linguistic values, enhancing the accuracy of fuzzy models in this way without loosing interpretability. This new type of fuzzy system can be optimized using suitable evolutionary algorithms. Drobics and Himmelbauer also emphasize the trade-off between between interpretability and accuracy in their model that is based on fuzzy regression trees. Instead of evolutionary algorithms they describe a three-stage optimization approach, defining first the underlying fuzzy sets taking the distribution of the data into account, then learning an initial fuzzy model that is fine-tuned and simplified in the final stage.

These first two papers are concerned with regression problems, where the aim is the prediction of a continuous variable. Wang, Nauck, Spott and Kruse propose to use fuzzy decision trees to solve classification problems, in which a nominal attribute has to be predicted. This fuzzy decision tree approach covers questions like the handling of missing values, pruning and the construction of the decision tree based on a generalized Shannon entropy. The paper by Serrurier and Prade is also concerned with supervised learning, however not with classification or regression problems in general, but with concept learning for which inductive logic programming is a very popular strategy. Serrurier and Prade introduce fuzzy rules into the framework of inductive logic programming in order to increase expressivity and adaptability.

The papers in the second group focus on exploratory data analysis and preprocessing, steps that are necessary before supervised learning can be applied. Nandi and Klawonn use different Takagi-Sugeno models in parallel for regression problems as a preprocessing step in order to detect ambiguities and inconsistencies in the data set that would make it impossible for general regression methods to fit the data in a meaningful way.

Cluster analysis is an exploratory data analysis technique to identify homogenous groups in multi-dimensional data sets. The evaluation of clustering results is often carried out on the basis of abstract validity measures that determine only a single number for the evaluation. The visualization of fuzzy clustering results introduced by Feil, Balasko and Abonyi is based on a modified multi-dimensional scaling technique and provides much more concise information on clustering results. Rehm, Klawonn and Kruse also use concepts from cluster analysis, however, not for the purpose

of grouping or partitioning the data, but for the identification of outliers, another important step in data preprocessing.

In closing this editorial, we would like to express our gratitude to the authors who contributed to a special issue that will definitely be of great interest for everyone working in the field of soft computing and information mining. Moreover, we would like to recognize the many reviewers who guaranteed the high quality of the papers that have finally been included.