Ersch. in: Intelligent Data Analysis ; 4 (2000), 2. - S. 93-95

Guest-editorial

Advances in intelligent data analysis

David J. Hand[a], Douglas H. Fisher[b] and Michael R. Berthold[c]
[a]*Department of Mathematics, Imperial College, 180 Queen's Gate, London, SW7 2BZ, UK*
E-mail: d.j.hand@ic.ac.uk; URL: http://www.ma.ic.ac.uk/statistics/djhand.html
[b]*Department of Computer Science, Box 1679, Station B, Vanderbilt University, Nashville, TN 37235, USA*
E-mail: dfisher@vuse.vanderbilt.edu; URL: http://cswww.vuse.vanderbilt.edu/~dfisher/
[c]*Berkeley Initiative in Soft Computing* (*BISC*), *Department of EECS, CS Division, University of California, Berkeley, CA 94720, USA*
E-mail: berthold@cs.berkeley.edu; URL: http://www.cs.berkeley.edu/~berthold

1. Introduction

Applications of Intelligent Data Analysis can be found in medicine, manufacturing, finance, agriculture, and many other areas of industry and science. In order to tackle the complicated and often enormous data sets which are being collected nowadays, methods from a diverse range of disciplines need to be combined. The interaction between these disciplines and the interaction with the user are open problems, meriting and stimulating a great deal of research.

To discuss these and related issues, Xiaohui Liu of Birkbeck College, University of London initiated a symposium series entitled "Intelligent Data Analysis", which started in 1995 at Baden-Baden, Germany [6]. After the successful second symposium in 1997 at London [1], a third symposium was held in 1999 at CWI in Amsterdam. Over 100 papers were submitted, of which an international program committee selected 45 for oral and poster presentation. A ballot was conducted throughout the conference and from the top-rated papers, five were chosen for this special issue. The authors of these five papers were invited to extend their conference paper and a second round of review ensured exceptional quality.

2. Content

This special issue combines five contributions from different backgrounds. Three methodological contributions are complemented by two papers describing application-oriented work that puts mixtures of different methodologies to work.

Many different kinds of data have to be analyzed and it is important that the method used incorporates as much a-priori information about the data as is available. The paper by Potharst and Bioch shows how a decision-tree-learning algorithm can use ordering information between the attributes, as well as ordering

information between the values of nominal attributes. The methodology presented generates monotone decision trees and allows non-monotone trees to be repaired. The resulting monotone decision trees are, in general, much smaller than the trees generated by the standard algorithm which ignores the ordering information.

It is obvious that the field of data analysis cannot develop in an application-free environment. To judge the importance of new developments (and mergers of existing ones), it is mandatory that new techniques be evaluated using real-world scenarios. Two contributions in this special issue address real world applications. In the paper by Flexer and Bauer techniques are discussed to find patterns in sequences of cognitive-evoked potentials in EEG recordings. They describe an unsupervised approach to finding such sequences and complement this technique by a visualization method. Huang and Zhao describe the analysis of large spatial data sets, in this case using weather features. Here a hierarchical approach extracts structures in weather data images using spatial aggregation. Through aggregation of adjacent features, a hierarchy of layers of such aggregations evolves, representing multiple intermediate representations.

During recent years the analysis of text documents has also received increasing attention. Not only are many more documents available in electronic form, but the emergence of the World Wide Web has provided us with a huge number of unstructured documents available online. So, in addition to classical document retrieval from databases it is becoming necessary to structure this online information. The paper by Hofmann describes an approach to visualizing the inherent organization of a collection of textual documents by topic. Through a statistical analysis, latent variables are extracted which model context-dependent word occurrences and form a topical decomposition of the documents. This information is then used to visualize the extracted topics in a two-dimensional map. Through different levels of granularity, a hierarchy of maps can be extracted, allowing the user to effectively browse through these topic maps in an interactive manner.

As was pointed out in [2] data analysis methods vary in the extent that they exploit a priori knowledge and "intelligence". It is easy to conduct unintelligent data analysis that wastes considerable time, only to yield uninformative results. Intelligent Data Analysis on the other hand, makes sure that the method(s) used will produce results that are useful, using computational power that is appropriate to the problem, and expressing the results in a form that is appropriate to the user. It is unlikely that all of these choices will be automated eventually, even though some work has been done in this direction (for example, on the semi-automatic selection of classifiers [5]). The paper by Mertens and Hand shows how boosting techniques can be adopted to automatically combine different types of classification surfaces. For full automation of model- and algorithm-selection, however, much more needs to be known.

3. Conclusions

If one compares the current state of Intelligent Data Analysis to the special issue that resulted from IDA-97 [7] progress in the area is obvious. More successful applications have emerged and more people have started to use methods from different disciplines. But even though IDA-99 was a stimulating and exciting event, much remains to be done. One of the most challenging issues for future conferences is likely to be the sheer size and enormous growth of the available data sets. These days data can be recorded so easily that the size of data sets often grow faster than computational power, outpacing Moore's famous law of semiconductor growth. This means that learning algorithms alone are not enough. Tools to explore these data sets interactively, breaking down the complexity and enabling the user to focus on aspects that are of current interest are also needed. Furthermore, algorithms that allow the user to access results

at any time throughout their execution are also of value. These enable first looks at rough structures which will become finer and more detailed once the user expends more computational power on the data analysis process [4]. In comparison to IDA-97, more contributions have focused on visualization of the analysis results, indicating that interactive visualization that allows the user to explore the structure in the data are a promising alternative to other knowledge extraction techniques. This is certainly an area of growing interest and we expect to hear more about related research at IDA-2001 as well.

Acknowledgements

We would like to thank all the authors for working with us on a very tight schedule. We are grateful to the reviewers who helped with the reviews despite our short notice: Elizabeth Bradley, Pavel Brazdil, Alois Heinz, Achim Hoffmann, Adele Howe, Ramon Lopez de Mantaras, Gokul Chander Prabhakar, Paola Sebastiani, Arno Siebes, and Tony C. Smith. Thanks also to Fazel Famili who made room for this special issue in his journal and also for all his help with editing this issue.

References

[1] M. Berthold, P. Cohen and X. Liu, Intelligent Data Analysis; Reasoning about Data, AAAi Press, *AI Magazine* **19**(4) (1998), 131–134.
[2] D.J. Hand, Intelligent Data Analysis, Issues and Opportunities, Elsevier Science Inc., *Intelligent Data Analysis* **2**(2) (1998), 67–79.
[3] D.J. Hand, J.J. Kok and M.R. Berthold, Advances in Intelligent Data Analysis, Lecture Notes in Computer Science (LNCS 1642), Springer Verlag, 1999.
[4] J.M. Hellerstein, R. Avnur, A. Chou, C. Hidber, C. Olston, V. Raman, T. Roth and P.J. Haas, Interactive Data Analysis: The Control Project, *IEEE Computer* (August 1999), 51–59.
[5] A. Kalousis and T. Theoharis, NOEMON: Design, implementation, and performance results of an intelligent assistant for classifier selection, Elsevier Science Inc, *Intelligent Data Analysis* **3**(5) (1999), 319–338.
[6] X. Liu, Intelligent Data Analysis: Issues and Challenges, *The Knowledge Engineering Review* **11**(4) (1996), 365–371.
[7] X. Liu, P. Cohen and M. Berthold, Editorial: Reasoning about Data, Elsevier Science Inc, *Intelligent Data Analysis* **2**(2) (1998), 63–66.