Intelligent Data Analysis 8 (2004) 437–438
IOS Press

Guest Editorial: "Advances in intelligent data analysis"

Michael R. Berthold[a], Elizabeth Bradley[b] and Rudolf Kruse[c]
[a]*University of Konstanz, Germany*
[b]*University of Boulder, Colorado, USA*
[c]*Otto-von-Guericke University, Magdeburg, Germany*

1. Introduction

Intelligent approaches to the analysis of large data sets continue to be of dramatic importance for many real world applications. Ranging from manufacturing, agriculture, finance and many other industrial and scientific areas, in the past few years increasingly large reservoirs of data of diverse type have been collected in the life sciences as well. In order to uncover the information hidden in these vast amounts of data, methods from different disciplines are required. Most prominently statistics and computer science but it has become increasingly clear that also detailed knowledge about the underlying domain is needed, especially in areas that attempt to analyse data from complex systems. The interaction between these disciplines with often very different vocabulary and the development of systems that interact with the user to find the desired answer(s) are still mostly open problems although in recent years encouraging progress has been made.

To discuss this and similar issues, Xiaohui Liu (now with Brunel University, UK) established a series of symposia, the first being held in Baden-Baden, Germany in the summer of 1995, followed by London (1997), Amsterdam (1999), and Lisbon (2001). In 2003 the fifth symposium was held in Berlin, Germany. Over 180 papers were submitted, of which an international program committee helped to select 17 for oral and 38 for poster presentation. Afterwards the Organizing Committee of the conference selected 6 papers for this special issue. The authors were asked to revise and extend their original contributions and an additional round of reviews resulted in the six papers presented here.

2. Content

This special issue contains six papers representing a nice cross section of the conference's program. The first three papers present more method oriented contributions, whereas the next two papers are application driven contributions. The last paper then shows one of many interesting new directions that work in the area of intelligent data analysis could also address.

The first part presents advances in cost-sensitive learning, conflicts in rule classification and modelling of changing time series. Geibel, Brefeld, and Wysotzki discuss the inclusion of misclassification costs into learning perceptron or SVM classifiers. Instead of focussing on class-dependent costs they allow exemplar-specific classification costs, which allows to adjust the cost function with higher accuracy. In their article on Resolving Rule Conflicts with Double Rule Induction, Lindgren and Boström present

a way to improve rule based classifiers by building specialized in areas of conflict. They propose to build a ne set of rules for pattern in the training set that are assigned to multiple classes. Experiments on benchmark data show substantial improvement over standard algorithms. The third contribution by Tucker and Liu presents a Bayesian approach to the modelling of changes in time series. The underlying Bayesian network is dynamic and the authors show how such a dynamic network structure can be learned.

Part two of this special issue presents two application driven methods, one coming from the text processing area and the other from biochemistry. Scheffer presents a method that allows to automatically answer a majority of emails for frequently asked questions. The article also discusses the benefit of co-training, that is the inclusion of unlabeled examples during the training process. The paper by Hofer, Borgelt, and Berthold presents an algorithm originally derived from the market basket analysis community to mine large molecular databases. The original algorithm had substantial performance problems and the approach presented here introduces a pre-processing step that identifies circular structures in all molecules considered, resulting in tremendous speed-ups and reduced memory usage. They also show a promising first attempt to mine frequent substructures that contain wildcard atoms, an important issue for biochemists.

The last paper, by Robins, Abernethy, Rooney, and Bradley, presents an essay on topology and intelligent data analysis. Shape is a deeply meaningful property in many different types of data. Sunspots, thunderstorms, and ice floes, for example, are coherent structures in their respective data sets, all of which are topologically conjugate to the unit circle. The authors argue that intelligent data analysis methods should make use of this universality.

3. Conclusions

The IDA symposium continues to be a stimulating and highly productive meeting. It is obvious when looking at the papers in the special issue and the overall program of the conference that the current state of the field is still very much in progress. Applications from more complicated areas such as for example bioinformatics continue to drive the development of optimized methods. We are still far away from a universal set of methods that allows to interactively explore enormous, diverse data and information sources while at the same time encouraging the user to incorporate her own expert knowledge in a truly intelligent way. We are looking forward to IDA-2005 to be held in Madrid, Portugal (August 22–25, 2005).

Acknowledgements

We would like to thank all of the authors for working with us on a rather tight schedule. Thanks also to Fazel Famili who once again, made room for this special issue in his journal.

Reference

[1] M.R. Berthold, H.-J. Lenz, E. Bradley, R. Kruse and C.H. Borgelt, eds, *Advances in Intelligent Data Analysis V, Lecture Notes in Computer Science*, LNCS 2810, Springer-Verlag, 2003.