Data and text mining

Data mining is a research area that concerns methods, techniques and tools for analyzing large data sets in order to support decision making. Research in this area may be found under several headings, including big data analytics, machine learning, knowledge discovery, predictive analytics and intelligent data analysis.

Text mining is a sub-field of data mining, which handles data in the form of natural language documents, e.g. news paper articles, web pages, scientific papers and electronic patient records.

Analysing large data sets to support decision making

A particular focus of our research is on predictive data mining using ensemble methods, i.e. techniques for generating sets of models that collectively form predictions by voting, and on methods for generating interpretable models, e.g. rule learning.

Our research on text mining focuses on efficient and resource lean methods using language technology for very large text sets. The research also focuses on semantic analysis, e.g. negation, speculation and temporality, in order to be able to extract situation-specific, accurate and relevant information from texts.

The main application area for our research is health care analytics, which aims for providing efficient and effective decision support for health care and pharmaceutical research.

The project *High-Performance Data Mining for Drug Effect Detection* is supported by the Foundation for Strategic Research with 19 MSEK during 2012-2016. The main goal of the project is to develop techniques and tools to support decision making and discovery of drug effects by analysing patient records, drug registries, case safety reports and chemical compound data in the form of both structured and unstructured (free text) data. The project will contribute with novel approaches to data mining and clinical text mining and develop a platform for large-scale analysis of massive, heterogeneous and continuously growing data sets.

The research group has collaborated for several years with computational chemists in the pharmaceutical industry. This has resulted in new techniques and tools for building predictive models from observed biological activities, e.g. toxicity, of chemical compounds, which are currently being used in the industry.

Another application area that we are involved in is modeling of component wear in heavy trucks using data mining and to provide decision support for optimizing heavy truck fleet utilization.

We participate in the project Integrated Dynamic Prognostic Maintenance Support (IRIS), which is lead by Scania AB, and supported by 11.6 MSEK from Swedish Governmental Agency for Innovation Systems (VINNOVA) during 2012-2017.

Contacts

Henrik Boström (data mining)

Lars Asker (data mining)

Hercules Dalianis (text mining)

Martin Duneld (text mining)

Tony Lindgren (data mining)

Panagiotis Papapetrou (data mining)

Sumithra Velupillai (text mining)

For contact information please consult www.dsv.su.se/datamining.

Focus areas

Ensemble methods

Interpretable models

Mining massive data sets

Resource lean text mining

Ongoing projects

High-Performance Data Mining for Drug Effect Detection

Integrated Dynamic Prognostic Maintenance Support

Department of Computer and Systems Science, DSV

www.dsv.su.se/datamining

Stockholm University