Chapter I

Automatic Intelligent Data Analysis

Martin Spott
Intelligent Systems Research Centre

Detlef Nauck
Intelligent Systems Research Centre

ABSTRACT

This chapter introduces a new way of using soft constraints for selecting data analysis methods that match certain user requirements. It presents a software platform for automatic data analysis that uses a fuzzy knowledge base for automatically selecting and executing data analysis methods. In order to support business users in running data analysis projects the analytical process must be automated as much as possible. The authors argue that previous approaches based on the formalisation of analytical processes were less successful because selecting and running analytical methods is very much an experience-led heuristic process. The authors show that a system based on a fuzzy knowledge base that stores heuristic expert knowledge about data analysis can successfully lead to automatic intelligent data analysis.

INTRODUCTION

Data is one of the most valuable assets of today’s businesses and timely and accurate analysis of available data is essential for making the right decisions and competing in today’s ever-changing business environment. Most businesses today face the analytics challenge and build ever-growing data warehouses but face a lack in analytical competence and resources. Data owners typically are domain experts who understand the processes that generated the data and what the data represents. However, data owners typically are not at the same time expert analysts and struggle with the
application of advanced analytics. Passing data on to an analyst results in a communication challenge that requires the domain expert to explain the data context and generate a problem statement that the analyst can use as the basis for analysing the data. When that has been done the analyst has to present the results in a way that the data owner can relate them to his context and derive valuable information for future decision making.

What then follows is an application challenge that requires IT support staff to turn analytical results into software that can be integrated into operational systems. With the introduction of data analysis functionality in databases and a standardised language for model descriptions like PMML (Predictive Model Markup Language) defined by the Data Mining Group (www.dmg.org), the integration may become simpler in the future. Under the assumption that the analysis tool is able to create a PMML description for the model in question and the database implements the underlying analysis algorithm, the PMML description can simply be included in a database script (e.g., PL/SQL for Oracle databases) that will be used to analyse data in the operational system. However, it still will take many years before data analysis is standard in databases and a large variety of models can be transferred in that way.

Commercial data analysis software that is aimed at a business context either is too simplistic and the manufacturer has decided to provide only limited functionality that non-expert users can handle or the software is too complex and provides advanced functionality that is aimed directly at expert analysts. In order to overcome both the analytics challenge and the communication challenge, tools are required that empower domain experts and data owners to run advanced analytics themselves with as little help from analysts as possible. One approach is to hide complex analytics under a layer of automation that provides an interface that allows users to work goal-oriented instead of method-oriented.

In this chapter we describe an approach to automating data analysis based on a fuzzy matching approach between user requirements and features of analytical models. We first discuss some general issues around automating analytics and then we present a software system that implements our approach.

**AUTOMATING DATA ANALYSIS**

When we talk about data analysis in this chapter we refer to the task of discovering a relationship between a number of attributes and representing this relationship in form of a model. Typically, we are interested in determining the value of some attributes given some other attributes (inference) or in finding groups of attribute-value combinations (segmentation). In this context we will not consider describing parameters of attribute distributions or visualisation.

Models are typically used to support a decision making process by inferring or predicting the (currently unknown) values of some output attributes given some input attributes or by determining a group to which the currently observed data record possibly belongs to. In this scenario we expect a model to be as accurate as possible. Models also can be used to explain a relationship between attributes. In this scenario we want a model to be interpretable.

A model is created in a (machine) learning process, where the parameters of the models are adapted based on set of training data. The learning process can be controlled by a separate validation set to prevent over-generalisation on the training set. The model performance is finally tested on a different test set.

In business environments data and problem owners are typically domain experts, but not data analysis experts. That means they do not have the required knowledge to decide which type of model and learning algorithm to choose, how to set the parameters of the learning procedure,