Available online at www.sciencedirect.com
SCIENCE DIRECT®
Applied Soft Computing 6 (2006) 348–356
Applied Soft Computing
www.elsevier.com/locate/asoc

Towards the automation of intelligent data analysis

Martin Spott *, Detlef Nauck

BT Research and Venturing, Intelligent Systems Research Centre, Orion pp1/12, Adastral Park, Ipswich IP5 3RE, UK

Abstract

Data analysis tools are still very much a collection of data analysis methods that require analysis experts as users. On the other hand, many business users are keen to apply data analysis to business data in order to understand it or to make predictions to improve on their business decisions. In order to make state-of-the-art data analysis techniques available to such non-experts, we developed a wizard for our data analysis tool SPIDA that selects appropriate data analysis methods given soft high-level requirements. The wizard also configures and runs the chosen methods automatically. This paper describes our general approach to automating data analysis, in particular, how to select an appropriate data analysis algorithm.

Keywords: Automatic data analysis; Fuzzy ranking; Computing with words

1. Introduction

Nowadays, data analysis tools are still very much a collection of data analysis methods that require analysis experts as users. The user not only needs domain knowledge of the data, but he also needs to know which data analysis methods are applicable to his problem, which ones meet some special requirements for the solution, how data needs to be prepared for the chosen method and finally, how the method needs to be configured.

On the other hand, many business users are keen to employ data analysis to make use of data that is being collected. Although a great proportion of typical analysis problems look quite simple to a data analysis expert, business users are overwhelmed by the sheer knowledge that is required to use current tools. We consider this one of the reasons for the fact that modern machine learning techniques like decision trees, neural networks, fuzzy techniques, support vector machines, etc. are still not industry standard.

Business users require a much more user- or problem-oriented approach to data analysis. Rather than knowing analysis methods, they are experts in the data domain and they know what they want to achieve with data analysis. If they only knew how. They might know, for example, that they want to classify insurance claims as fraudulent or non-fraudulent, given historic information of the customer and the current case. They might want to understand, how the analysis method actually classifies customers (e.g. with a rule set), they might require a certain classification accuracy and that the algorithm is so simple that it can be implemented as an SQL query. Ideally, such users would simply like

* Corresponding author.
E-mail addresses: martin.spott@bt.com (M. Spott), detlef.nauck@bt.com (D. Nauck).

1568-4946/$ – see front matter
doi:10.1016/j.asoc.2005.11.002

to feed all these high-level requirements and the data into a tool that would then automatically find the best algorithm in terms of requirements, configure it, run it and create a software module that can be plugged into the business application.

Based on these ideas we developed SPIDA (soft computing platform for intelligent data analysis) [7] and equipped it with a wizard that, to certain extent, does most of the things mentioned above. The rest of the paper is organised as follows.

In Section 2, we discuss the general problem of automating data analysis and existing approaches. Afterwards, we show how to use soft constraints to rank data analysis methods in Section 3. Finally, we describe the wizard of our data analysis tool SPIDA as an implementation of the techniques introduced before.

2. Approaches to automatic data analysis

A typical data analysis tool provides access to data sources like databases or plain text files, it can filter the data in various ways and prepare it for the actual data analysis, for example change format or representation. Such a tool offers a variety of data analysis methods which can be applied to prepared data and can finally present results in different formats. Setting up a data analysis process involves the following steps.

(1) Define the data analysis problem (like classification, prediction, clustering).
(2) Define requirements and preferences for the solution.
(3) Select the data source.
(4) Depending on the data and the problem.
    (a) Decide on applicable data analysis methods.
    (b) Filter and prepare data according to chosen methods (might be different for different methods).
    (c) Configure data analysis methods (parameter settings).
(5) Run the analysis.
(6) Check the results.
(7) Go back to step 4a if results are not satisfactory.
(8) Produce report.

The first two steps define what the analyst wants to achieve, i.e. what he wants to solve and how he wants the solution to look like. From the business perspective, this should include how the solution will be applied like if it will be a stand-alone application, a module embedded in an existing application, a simple SQL query, etc. Depending on such requirements, some analysis methods are more suitable than others. From step 4a on typically a data analysis expert is required, who knows from experience which data analysis methods can be applied to which problem given a number of requirements and the data. The expert would also know, what kind of data preparation is necessary and how to set the parameters of the chosen methods. After setting up the analysis process it will be run and the results will be evaluated. Usually, first results are not satisfactory, so the actual data analysis process is iterative. Based again on experience, an expert might change data preparation and parameter settings in the actual analysis method several times in order to achieve better results.

Our understanding of automatic data analysis incorporates steps 4–8. The user is required to define the type of the problem, his requirements and preferences for the solution as well as specify the data source. Given this information, an automated tool would take over and recommend appropriate analysis methods, set-up the analysis process, run it, match the results with the given requirements and improve the match iteratively by changing the set-up, if necessary.

In the following subsection we briefly describe existing approaches for automating data analysis. Our approach will be introduced in subsequent sections, whereby we focus on selecting the most suitable data analysis method according to user requirements.

2.1. Previous approaches

The approach described in [11] breaks analysis methods down into formal blocks and represents user requirements in a formal language. Then a search algorithm identifies suitable blocks and arranges them in a way to carry out an analysis process. This approaches faces the problem of formalising mainly heuristic methods and that it is usually not feasible to formally compute all necessary parameters to execute an analysis method.

Other authors discuss mainly architectural features of systems that could automate data analysis or data