Computers & Graphics 41 (2014) 99–113

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on Uncertainty and Parameter Space Analysis in Visualization

Interactive exploration of parameter space in data mining: Comprehending the predictive quality of large decision tree collections

Luciana Padua [a], Hendrik Schulze [b], Krešimir Matković [b,*], Claudio Delrieux [c]

[a] Universidad de Buenos Aires, Argentina
[b] VRVis Research Center in Vienna, Austria
[c] Universidad Nacional del Sur, Bahia Blanca, Argentina

ARTICLE INFO

Article history:
Received 15 July 2013
Received in revised form
8 February 2014
Accepted 26 February 2014
Available online 18 March 2014

Keywords:
Decision trees
Parameter space exploration
Visual analytics
Knowledge discovery

ABSTRACT

Decision trees are an intuitive yet powerful tool for performing predictive data analysis in data mining. In order to generate an adequate predictive model from a data set, a data analyst has to assess the predictive quality of the decision trees derived from several combinations of working parameters. Except in very simple cases, this may be a tedious and error prone supervised task, since the parameter space is frequently huge. Analysts rely on their intuition and usually test just a few different parameter settings. In this work we present an interactive approach to facilitate the comprehension of the predictive power of large collections of decision trees by exploring large portions of the parameter space. For this, we developed novel views that allow us to visualize and analyze the predictive quality of hundreds of trees, working together with coordinated multiple views of tree representations (needed to understand the tree shapes and actual information herein), and aggregates of Receiver Operating Characteristic (ROC) and lift curves for assessing the predictive quality of the models. We developed a worked example using a data set from a Telecommunications company, showing how easy and natural it is to gain insight into the behavior of the data within our exploration tool, as compared with the traditional and widespread common practice of data analysts.

1. Introduction

Data analysis, as part of the knowledge discovery in databases (KDD) process, is aimed to build useful models from raw data. These models, in turn, produce predictive results that may be useful in a wide spectrum of specific purposes. In the verge of the Big Data revolution, the requests for interactive, real-time data analysis are becoming commonplace in areas that may range from DNA analysis to business strategies, scientific discovery, recommendation systems, textual corpus understanding, and econometric analysis, to mention just a few. In particular, classification is an omnipresent task in data analysis in practically all domains. Data analysts want to classify information items based on a specific model, with which they may use the knowledge of prior examples or cases to predict outcomes or behaviors in newer situations. Predictive analysis is the specific task of building predictive models. It has been one of the most active research areas in data analysis and machine learning during the last few decades. A predictive model requires to identify the relevant explanatory variables and their specific interplay, in a way such to have sufficient conditions (backed on prior examples) to assess the likelihood of a particular outcome in a given situation.

Even though there is a large set of predictive model types, decision tree learning is still the most popular and better understood since its inception several decades ago. Tree learning is a method for approximating discrete valued target functions, in which the learned function is represented by a tree [1]. Trees, as a support tool, represent all the different outcomes or possibilities of a given decision process. If the outcome is a categorical target, then the tree is called a *decision tree* (DT). Instead, if the goal is to predict a continuous variable, the tree is referred to as a *regression tree*. DTs are produced by algorithms that identify various ways of splitting a data set into branch-like segments [2]. This procedure generates an inverted tree that starts with a root node containing all the records of the training data set and ends with several terminal nodes, also called leaves. In order to build the DT, at each node a splitting decision is analyzed and contrasted to a set of learning parameters. The goal is to look for relationships between the input data and the target value. Each leaf represents a rule that classifies the data into a class. Once the building process ends, the decision rules generated can be used to predict the classes of the remaining records in the data set. To measure the predictive power of these rules many measures are calculated, being the most used ones specificity, sensitivity, recall, ROC, and lift curves.

* Corresponding author.

http://dx.doi.org/10.1016/j.cag.2014.02.004
0097-8493/

The whole process of finding an adequate DT is of exponential complexity, and therefore an exploration of the whole set of possible DTs for a given example set is out of the question in interactive analysis. Even applying the DT learning techniques proposed in the machine learning community is not adequate if the data set is reasonably large.

For this reason, it is common practice in KDD, for instance in business intelligence, to develop tools that facilitate the work of data analysts in the manipulation of a set of parameters to produce different models and results. The current practice among analysts is to progress by trial-and-error, which relies strongly on their intuition and experience. In addition, the relation between the parameter space and the mining results is usually not apparent or able to be displayed in a visually understandable way. In other words, the relationship between the selected parameter values and the mining results can be considered as a kind of black box.

Information visualization can be a remarkable aid in this process. Visualization has been used mainly in two stages of the KDD process [3]. At the beginning, to analyze and reduce the data, and at the end to display the results of the trained model. However, not much work has been done on the intermediate, more essential steps of KDD, namely the actual learning or discovery process. In this paper we present a novel approach for exploring the predictive power of large collections of decision trees. We use InfoVis techniques (a) to represent the relationship between parameters and results, (b) to facilitate the comparison of alternative results, and (c) to enable an overall comprehension of the properties of the (usually very large) DT collection generated during the exploration. For these purposes, we designed two new views, the Cross-Parameter View which depicts complex correlations between parameters, and the Multiple Trees Explorer View, to navigate through the whole tree collection. Both views, together with other traditional InfoVis views, are integrated in a coordinated multiple views system.

We illustrate the results with a worked example using a real-word data set from a Telecommunications company, where the purpose is to develop predictive models of land-line customers' behavior. We show several examples of the discovery of valuable, though not obvious, features of the predictive quality of collections of DTs generated under different parameter combinations. This insight was gained with the straightforward use of our exploration tool, exhibiting a clear advantage over the usual practice of data analysts.

2. Related work

The areas of Visualization and of KDD have developed independently. However, recently many efforts have been done to integrate both domains, and a research community has been devoted to work on what is known today as Visual Analytics. According to [3], this new approach is about "*combining automated analysis techniques with interactive visualizations for effective understanding, reasoning and decision making on the basis of very large complex data sets*". Bertini and Lalanne [4] believe that Visualization's contribution to KDD could be enhanced in two ways, first by providing means to more directly represent the relationship between parameters and results, and second, by allowing visualization structures that enable the comparison of alternative results. In particular, the quest for interactive visual analysis in parameter space exploration is not new. It has been used to explore parameter spaces in different domains, such as fishery [5] or engineering [6–8], for example. In the KDD specific domain we can reference work done for clustering parameters [9] and neural network parameters [10]. In general, these approaches to parameter set exploration appear to cope very barely with the exponential nature of the parameter space, and with the needs of a data analyst requiring to keep track of the results in a KDD context.

On the other hand, tree visualization is an active area of research [11]. Visualizing or analyzing interactively a large set of trees is quite a challenging problem. Munzner et al. [12] proposed an interactive method to compare two, very large, trees. Bremm et al. [13] introduced an interactive way to compare several trees. Their approach is tailored to trees from the biology domain, and they do not try to support the trees creation process. Graham [14] provided a very comprehensive survey on multiple trees visualization. Finally, Van den Elzen and van Wijk [15] introduced the BaobabView. Their motivation is to support the tree generation process, gradually building a tree and exploring it on the way. Again, these results appear to be of little help if tree visualization is used for result comparison in a mining application.

Apart from interactive visual exploration of the parameter space, and tree visualization, the KDD process can also benefit from support tools that aid the data analyst to keep track and coordinate different aspects and views that result from the mining process. The visual generation of support tools from raw data closes the gap imposed by the required levels of abstraction of the visualization pipeline. However, the application of sophisticated Visual Analytics to classification problems is still a relatively open field. May and Kohlhammer [16] proposed a methodology for coupling data classification and interactive visualization that makes all the abstraction levels visible and steerable. Afzal et al. [17] developed tools for interactive decision support for exploring infectious diseases. An analysis of the complex behavior of computer network systems is presented by Teoh et al. [18], where the authors developed a system for detecting flaws and intruders analyzing the log files.

In general, these techniques and proposals, by themselves or together, are still insufficient for a data analyst faced with the burden of exploring the parameter space of a reasonably large data set (tens of thousands of records or larger) to find useful predictive models. No relevant contribution appears to have been presented that takes advantage of visual support tools in the specific application domain of interactive exploration of parameter space for DT learning, and thus the motivation for the following development.

3. Decision tree generation and assessing in KDD

As was mentioned above, DTs are a key knowledge representation feature in the KDD process. In a DT, each node is labeled with a specific test or decision on an attribute, and the successors or outbound branches from the node represent the possible outcomes of that test. The leaf nodes, instead, are not labeled with tests but with outcome classes, representing the obtained outcome (given the specific sequence of tests or decisions taken from the root of the tree to a given leaf). DTs are quite popular due to their powerful predictive capabilities, their simplicity to create classification rules, and the fact that they represent knowledge in a crisp and intuitive fashion (for instance, they can easily be translated into everyday language). Just as a simple example, suppose that we have a choice of three different activities on Saturday evenings: cinema, opera, visiting a pub. We visit a pub in summer, in winter we go the opera if Nabucco is playing, otherwise we go to the cinema. A decision tree which describes our activities is shown in Fig. 1.

Tree construction proceeds top-down. A test on the value of a given attribute is applied to the root. The possible arising outcomes split the set of cases in their corresponding successor nodes, and the associated branches are tagged with their respective $<\texttt{attribute}=\texttt{value}>$ pairs. The new nodes thus obtained