Computational Statistics and Data Analysis 71 (2014) 986–1000

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

(Psycho-)analysis of benchmark experiments: A formal framework for investigating the relationship between data sets and learning algorithms

Manuel J.A. Eugster [a,*], Friedrich Leisch [b], Carolin Strobl [c]

[a] Institut für Statistik, Ludwig-Maximilians-Universität München, Ludwigstrasse 33, 80539 Munich, Germany

[b] Institute of Applied Statistics and Computing, University of Natural Resources and Life Sciences, Vienna, Gregor Mendel-Strasse 33, 1180 Vienna, Austria

[c] Department of Psychology, University of Zurich, Binzmühlestrasse 14, 8050 Zurich, Switzerland

ARTICLE INFO

Article history:
Received 16 March 2012
Received in revised form 31 July 2013
Accepted 7 August 2013
Available online 16 August 2013

Keywords:
Benchmark experiments
Data set characterization
Recursive partitioning
Preference scaling
Bradley–Terry model

ABSTRACT

It is common knowledge that the performance of different learning algorithms depends on certain characteristics of the data—such as dimensionality, linear separability or sample size. However, formally investigating this relationship in an objective and reproducible way is not trivial. A new formal framework for describing the relationship between data set characteristics and the performance of different learning algorithms is proposed. The framework combines the advantages of benchmark experiments with the formal description of data set characteristics by means of statistical and information-theoretic measures and with the recursive partitioning of Bradley–Terry models for comparing the algorithms' performances. The formal aspects of each component are introduced and illustrated by means of an artificial example. Its real-world usage is demonstrated with an application example consisting of thirteen widely-used data sets and six common learning algorithms. The Appendix provides information on the implementation and the usage of the framework within the R language.

1. Introduction

The aim of this paper is to introduce a new formal framework for describing the relationship between data set characteristics and the performance of learning algorithms. The main target audience for employing this framework is methodologists from statistics and machine learning. A major part of their research consists of developing new and better learning algorithms and, consequently, trying to investigate on which types of problems their new algorithms outperform existing ones (either by means of simulated or by means of real data sets). In the development of new statistical methods, it is a common course of action that the construction of a new algorithm is guided by the poor performance of existing methods on a particular type of problem—such as high-dimensional or not linearly separable data. However, it is by no means trivial to document the superiority of an algorithm for certain kinds of data sets in an objective and statistically sound way. To address this issue, the approach presented in the following offers a reliable framework for illustrating the properties of a statistical learning algorithm in a benchmark study. This applies to the task of developing new and better learning algorithms

* Corresponding author. Tel.: +49 89 2180 6254; fax: +49 89 2180 5040.
E-mail addresses: Manuel.Eugster@stat.uni-muenchen.de (M.J.A. Eugster), Friedrich.Leisch@boku.ac.at (F. Leisch), Carolin.Strobl@psychologie.uzh.ch (C. Strobl).
URLs: http://www.statistik.lmu.de/~eugster (M.J.A. Eugster), http://www.rali.boku.ac.at/friedrichleisch.html (F. Leisch), http://www.psychologie.uzh.ch/fachrichtungen/methoden/team/carolinstrobl.html (C. Strobl).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.08.007

as well as to the task of comparing the performance of a set of established ones (either to illustrate their known properties or to discover which algorithms perform particularly well in certain situations).

The presented framework combines the advantages of three well-established approaches: *Benchmark experiments* from statistical and machine learning to evaluate the performance of the algorithms; *statistical and information-theoretic measures* from meta-learning to describe the data sets; and recursive partitioning of *Bradley–Terry models* from psychology to capture the differences in the performance of the algorithms on data sets with similar characteristics. It is intended to be a practical tool for exploring and learning about the relationship between certain data set characteristics and learning algorithms. On this account, we also provide an easy to use implementation of the framework in the R programming language (R Development Core Team, 2011).

The article is organized as follows: Section 2 motivates the three components of our framework and positions the framework within related work. Section 3 presents all related methods in detail. First, we introduce a formal notation for benchmark experiments; then we define a sound and flexible framework for data set characterization and introduce a common set of data set characteristics; finally, we outline the principle of model-based recursive partitioning and its generalization to Bradley–Terry models. A toy example is used to demonstrate each part of the framework. Section 4 then applies the proposed method to a real-world example based on classification problems from the well-known UCI Machine Learning Repository. The article is concluded with a summary and an outlook in Section 5. Appendix A demonstrates the implementation of the framework in the R language (R Development Core Team, 2011). There we replicate the toy example that can serve as a template for applying the framework to the readers' own data sets and algorithms. Finally, Appendix B provides general computational details for replicating the article.

2. Related work and motivation

In contrast to the well-known idea of meta-learning (see, e.g., Vilalta and Drissi, 2002) the presented approach follows a radically different idea: the transformation of meta-learning from a prediction problem to an analysis problem. Instead of providing a recommender system for algorithms similar to the recommender systems for movies or books (a common problem in machine learning), here the aim is to provide a systematic, objective, reproducible, and statistically sound way to investigate the relationship between certain data set characteristics and learning algorithms for a given problem domain.

Usually, performance is investigated on a collection of data sets, e.g., from the UCI Machine Learning Repository (Asuncion and Newman, 2007). It is well known that the characteristics of the data sets have an influence on the performance of the algorithms—almost every publication that proposes a new algorithm presents its performance on data sets with certain characteristics (even though often only the number of observations and attributes vary). Nonetheless, in most publications differences of the data sets are noted but not used for further analyses (perhaps the best known study is STATLOG by King et al., 1995, newer ones are, e.g., Lim et al., 2000 and Caruana et al., 2008). An approach incorporating both algorithms and data sets was suggested by Kalousis et al. (2004), who investigate the relations between learning algorithms and data sets by means of clustering the algorithms on one hand and the data sets on the other hand based on the performance measures. These cluster results (a large number of graphics) are then visually interpreted by a human decision maker to find relations. The present article is an enhancement of this unsupervised and subjective approach, and provides an automated framework where each step of the relation finding process is based on sound statistical methodology.

In statistical and machine learning, benchmark experiments are empirical investigations with the aim of comparing and ranking learning algorithms with respect to certain performance measures (see, e.g., Torti et al., 2012; Givens et al., 2013). We propagate the benchmark experiments based on bootstrapping or subsampling without summarizing the results (e.g., computing the mean). To our best knowledge, existing approaches use k-fold cross-validation to estimate the algorithms' performances and then compute a summary measure based on the performances estimated in each fold. This reduction to one summary statistic discards a lot of information. In our approach, the different samples from the data sets or experimental settings resemble individual subjects that "vote" on the performance of the algorithms. By not summarizing the results before this analysis, the approach takes into account the sampling variability and naturally incorporates concepts like ties.

In psychology and related disciplines, the pairwise comparative choice model suggested by Bradley and Terry (1952) is the most widely used method to study preferences of *subjects* (e.g., consumers or patients) on some *objects* (e.g., a set of chocolate bars or different pain therapies). The preference scaling of a group of subjects may not be homogeneous, but different groups of subjects with certain characteristics may show different preference scalings. A newly developed semi-parametric approach for recursive partitioning of Bradley–Terry models (Strobl et al., 2011) takes this circumstance into account: It identifies groups of subjects with homogeneous preference scalings in a data-driven and statistically sound way. This approach is an extension of the classical algorithms for classification and regression trees (CART Breiman et al., 1984; Quinlan, 1993), but avoids the statistical flaws of these early algorithms (Hothorn et al., 2006; Strobl et al., 2009). It results in a tree where the subjects are divided into groups according to their characteristics, and in each terminal leaf a Bradley–Terry model shows the preference scaling within this group.

The use of Bradley–Terry models has already been suggested for deriving consensus rankings from benchmark studies (Hornik and Meyer, 2007). However, in order to utilize the information inherent in different characteristics of the data sets, here we suggest to apply the advanced approach of recursive partitioning of Bradley–Terry models in the analysis of benchmark studies. In this framework, the data sets are the subjects and the algorithms are the objects. The interpretation is