Available online at www.sciencedirect.com

SCIENCE DIRECT®

European Journal of Operational Research 173 (2006) 781–800

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/ejor

The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing

Sven F. Crone [a], Stefan Lessmann [b,*], Robert Stahlbock [b]

[a] Department of Management Science, Lancaster University, Lancaster LA1 4YX, United Kingdom

[b] Institute of Information Systems, University of Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany

Received 15 November 2004; accepted 18 July 2005
Available online 15 November 2005

Abstract

Corporate data mining faces the challenge of systematic knowledge discovery in large data streams to support managerial decision making. While research in operations research, direct marketing and machine learning focuses on the analysis and design of data mining algorithms, the interaction of data mining with the preceding phase of data preprocessing has not been investigated in detail. This paper investigates the influence of different preprocessing techniques of attribute scaling, sampling, coding of categorical as well as coding of continuous attributes on the classifier performance of decision trees, neural networks and support vector machines. The impact of different preprocessing choices is assessed on a real world dataset from direct marketing using a multifactorial analysis of variance on various performance metrics and method parameterisations. Our case-based analysis provides empirical evidence that data preprocessing has a significant impact on predictive accuracy, with certain schemes proving inferior to competitive approaches. In addition, it is found that (1) selected methods prove almost as sensitive to different data representations as to method parameterisations, indicating the potential for increased performance through effective preprocessing; (2) the impact of preprocessing schemes varies by method, indicating different 'best practice' setups to facilitate superior results of a particular method; (3) algorithmic sensitivity towards preprocessing is consequently an important criterion in method evaluation and selection which needs to be considered together with traditional metrics of predictive power and computational efficiency in predictive data mining.

Keywords: Data mining; Neural networks; Data preprocessing; Classification; Marketing

1. Introduction

In competitive consumer markets, data mining faces the growing challenge of systematic knowledge discovery in large datasets to achieve

[*] Corresponding author. Tel.: +49 40 42838 5500; fax: +49 40 42838 5535.

E-mail addresses: s.crone@lancaster.ac.uk (S.F. Crone), lessmann@econ.uni-hamburg.de (S. Lessmann), stahlboc@econ.uni-hamburg.de (R. Stahlbock).

0377-2217/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2005.07.023

operational, tactical and strategic competitive advantages. As a consequence, the support of corporate decision making through data mining has received increasing interest and importance in operational research and industry. As an example, direct marketing campaigns aiming to sell products by means of catalogues or mail offers [1] are restricted to contacting a certain number of customers due to budget constraints. The objective of data mining is to select the customer subset most likely to respond in a mailing campaign, predicting the occurrence or probability of purchase incident, purchase amount or interpurchase time for each customer [2,3] based upon observable customer attributes of varying scale. Traditionally, response modelling has utilised transactional data consisting of continues variables to predict purchase incident focusing on the recency of the last purchase, the frequency of purchases and the overall monetary purchase amount, referred to as recency, frequency and monetary value (RFM)-analysis [2]. The continuous scale of these attributes together with their limited number has facilitated the use of conventional statistical methods, such as logistic regression.

Recently, progress in computational and storage capacity has enabled the accumulation of ordinal, nominal, binary and unary demographic and psychographic customer centric data, inducing large, rich datasets of heterogeneous scales. On the one hand, this has advanced the application of data driven methods like decision trees (DT) [4], artificial neural networks (NN) [2,5,6], and support vector machines (SVM) [7], capable of mining large datasets. On the other hand, the enhanced data has created particular challenges in transforming attributes of different scales into a mathematically feasible and computationally suitable format. Essentially, each customer attribute may require special treatment for each algorithm, such as discretisation of numerical features, rescaling of ordinal features and encoding of categorical ones. Applying a variety of different methods, the phase of data preprocessing (DPP) represents a complex prerequisite for data mining in the process of knowledge discovery in databases [8].

Aiming to maximise the predictive accuracy of data mining, research in management science and machine learning is largely devoted to enhancing competing classifiers and the effective tuning of algorithm parameters. Classification algorithms are routinely tested in extensive benchmark experiments, evaluating the impact on predictive accuracy and computational efficiency, using preprocessed datasets; e.g. [9–11]. In contrast to this, research in DPP focuses on the development of algorithms for particular DPP tasks. While feature selection [12–14], resampling [15,16] and the discretisation of continuous attributes [17,18] are analysed in some detail, few publications investigate the impact of data projection for categorical attributes and scaling [19,20]. More importantly, interactions on predictive accuracy in data mining are not been analysed in detail, especially not within the domain of corporate direct marketing.

To narrow this gap in research and practice, we seek to investigate the potential of DPP in a real world scenario of response modelling, predicting purchase incident to identify those customers most likely to respond to a mailing campaign in the publishing industry. We analyse the impact of different DPP schemes across a selection of established data mining methods. Due to the questionable usefulness of traditional statistical techniques in large scale data mining settings [21,22] and mixed scaling levels of customer attributes, we confine our analysis to data driven methods of C4.5 DT, NN and SVM.

The remainder of the paper is organised as follows: We begin with a short overview of the classification methods of DT, NN and SVM used. Next, the task of DPP for competing methods for scaling, sampling and coding is discussed in Section 3. Conducting a structured literature review, we exemplify that the influence of DPP is widely overlooked to motivate our further analysis. This is followed by the case study setup of purchase incident modelling for direct marketing in Section 4 and the experimental results providing empirical evidence for the significant impact of DPP on classification performance in Section 5. Conclusions are given in Section 6.