Chapter XI
Effective Intelligent Data Mining Using Dempster–Shafer Theory

Malcolm J. Beynon
Cardiff University, UK

ABSTRACT

The efficacy of data mining lies in its ability to identify relationships amongst data. This chapter investigates that constraining this efficacy is the quality of the data analysed, including whether the data is imprecise or in the worst case incomplete. Through the description of Dempster-Shafer theory (DST), a general methodology based on uncertain reasoning, it argues that traditional data mining techniques are not structured to handle such imperfect data, instead requiring the external management of missing values, and so forth. One DST based technique is classification and ranking belief simplex (CaRBS), which allows intelligent data mining through the acceptance of missing values in the data analysed, considering them a factor of ignorance, and not requiring their external management. Results presented here, using CaRBS and a number of simplex plots, show the effect of managing and not managing of imperfect data.

INTRODUCTION

The considered generality of the term data mining highlights the wide range of real-world applications that have benefited or could benefit from its attention. It also acknowledges the increasing amount of information (data) available when insights into a problem are the intent. Its remit certainly encompasses the notion of secondary data analysis, as referred to in a definition of data mining given in Hand (1998, p. 112), who state: "*the process of secondary analysis of large databases aimed at finding unsuspected relationships which are of interest or value to the database owners*."

However, at the start of the 21st century, it could be viewed data mining has matured beyond this, to take in primary data analysis also, where data is

Copyright © 2007, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.

collected and analysed with a particular question or questions in mind (Hand, 1998). Peacock (1998), highlight the confusion to what is data mining, suggesting it is defined within a narrow scope by some experts, within a broad scope by others, and within a very broad scope by still others. A reason for this confusion is because of the evolution in the range of data mining techniques available to an analyst, many benchmarked using more primary data where results are prior known.

One direction of this evolution is within the environment of uncertain reasoning, which by its definition acknowledges the often imperfection of the considered data. Chen (2001), outlines the rudiments of uncertain reasoning based data mining, highlighting the often reality of the presence of imprecision and incompleteness of the available data to be analysed. Amongst the associated general methodologies considered, including rough set theory (Pawlak, 1982) and fuzzy set theory (Zadeh, 1965), is Dempster-Shafer theory, introduced in Dempster (1967, 1968) and Shafer (1976). DST is often described as a generalisation of the well-known Bayesian theory (Shafer & Srivastava, 1990), noticeably further developed in the form of the transferable belief model (Smets, 1990; Smets & Kennes, 1994). Inherent with DST is its close association with the ability to undertake data mining in the presence of ignorance (Safranek, Gottschlich, & Kak, 1990).

A nascent DST-based technique for object classification and ranking is the classification and ranking belief simplex. Introduced in Beynon (2005a), it encompasses many of the advantages the utilisation of DST can bestow on knowledge discovery and data mining. The advantages highlighted here are central to effective data mining, including; the non-requirement for knowledge on specific data distributions, the ability to work with the presence of missing values without the need for their inhibiting management and the mining of low quality information. CaRBS further offers a visual representation of the contribution of data to the classification of objects using simplex plots, including the concomitant levels of ambiguity and ignorance (Beynon, 2005b).

The main emphasis in this chapter is on the presence of missing values in data and its effect on the subsequent data mining. A need to concern oneself with this issue is that most data mining techniques were not designed for their presence (Schafer & Graham, 2002). The reality is however that there exists a nonchalant attitude to their presence. Exemplified in Barnett (2005), when describing data mining within customer information systems, they comment that, "it is easier to produce accurate reports even when data is missing—through statistical adjustment." It is as though simply managing (replacing) missing values solves the problem, without realising that it brings its own disadvantages, namely a different dataset to that originally available. The CaRBS technique does not require any management of missing values, instead considering them as concomitant ignorance, so allowing data mining on the richness of the original data. This effective utilisation of belief functions (a more general term for DST), offers one direction for the mitigation of the comment in Zaffalon (2002, p. 108), who suggests: "statistical treatment of missing data does not appear to have benefited yet from belief function models."

A "complete" bank rating dataset (with no missing values) is initially analysed using the CaRBS technique. Moreover, a simplified binary classification of the Fitch bank individual rating (FBR) on large U.S. banks is considered, with each bank described by a number of financial variables (FitchRatings, 2003). Similar CaRBS analyses are undertaken when a large proportion of the financial values are denoted missing, with comparable results presented when the missing values are retained and when they are managed through imputation (see Huisman, 2000).

The relative simplicity of the CaRBS technique and visual presentation of findings allows the reader the opportunity to succinctly view a

19 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/effective-intelligent-data-mining-using/24908

Related Content

Replacing Support in Association Rule Mining
Rosa Meoand Dino Ienco (2010). *Rare Association Rule Mining and Knowledge Discovery: Technologies for Infrequent and Critical Event Detection (pp. 33-46).*
www.irma-international.org/chapter/replacing-support-association-rule-mining/36898

A Comparative Study of Associative Classifiers in Mesenchymal Stem Cell Differentiation Analysis
Weiqi Wang, Yanbo J. Wang, Qin Xin, René Bañares-Alcántara, Frans Coenenand Zhanfeng Cui (2011). *Knowledge Discovery Practices and Emerging Applications of Data Mining: Trends and New Domains (pp. 223-243).*
www.irma-international.org/chapter/comparative-study-associative-classifiers-mesenchymal/46898

Boosting Prediction Accuracy of Bad Payments in Financial Credit Applications
Russel Pearsand Raymond Oetama (2010). *Rare Association Rule Mining and Knowledge Discovery: Technologies for Infrequent and Critical Event Detection (pp. 255-269).*
www.irma-international.org/chapter/boosting-prediction-accuracy-bad-payments/36911

Communication Matrices for Managing Dialogue Change to Teamwork Transformation
James Calvin (2020). *Optimizing Data and New Methods for Efficient Knowledge Discovery and Information Resources Management: Emerging Research and Opportunities (pp. 98-115).*
www.irma-international.org/chapter/communication-matrices-for-managing-dialogue-change-to-teamwork-transformation/255753

Conceptual Approach to Predict Loan Defaults Using Decision Trees
Syed Muzamil Basha, Dharmendra Singh Rajputand N. Ch. S. N. Iyengar (2019). *Sentiment Analysis and Knowledge Discovery in Contemporary Business (pp. 148-161).*
www.irma-international.org/chapter/conceptual-approach-to-predict-loan-defaults-using-decision-trees/210968