Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

ELSEVIER International Journal of Approximate Reasoning 35 (2004) 29–53

INTERNATIONAL JOURNAL OF APPROXIMATE REASONING

www.elsevier.com/locate/ijar

Stability of continuous value discretisation: an application within rough set theory

Malcolm J. Beynon *

Cardiff Business School, Cardiff University, Colum Drive, Cardiff CF10 3EU, Wales, UK

Received 1 November 2002; received in revised form 1 April 2003; accepted 1 May 2003

Abstract

Continuous value discretisation (CVD) is the process of partitioning a set of continuous values into a finite number of intervals (categories). This paper introduces a number of stability measures associated with the resultant CVD. The stability measures are constructed from a series of estimated probability distributions for the individual 'partitioning' intervals found using the method of Parzen windows. These measures enable comparisons between the results of alternative methods of CVD on their ability to effectively partition the continuous values. A further utilisation of these measures is exposited within rough set theory (RST). RST is a modern approach to the generation of sets of rules enabling the classification of objects to categories based on sets (reducts) of related characteristics. To avoid rules of poor quality (from RST analysis) induced directly from continuous valued characteristics, CVD methods can be used to reduce the associated granularity and allow higher rule quality. The notion of stability introduced enables the further introduction of novel measures particular to reduct and rule set stability within RST.

Keywords: Data discretisation; Parzen windows; Stability; Rough set theory

1. Introduction

Continuous value discretisation (CVD) is the process of partitioning a set of continuous values (data) into a finite number of intervals (categories). A simple

* Tel.: +44-29-2087-5747; fax: +44-29-2087-4419.
E-mail address: beynonmj@cardiff.ac.uk (M.J. Beynon).

0888-613X/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S0888-613X(03)00057-4

example of CVD is the categorisation of continuous values into a given number of intervals based on equidistant cut-points (equal width discretisation). The study of CVD is an ongoing research topic, including specific CVD technique development and comparison between techniques [20,29]. A necessity for data to be discretised may be to improve the utilisation of certain symbolic machine learning methods, including rough set theory (RST), which is a rule based technique for object classification. In the case of RST, to avoid rules of poor quality induced directly from continuous valued characteristics, CVD techniques can be used to reduce the associated granularity and allow higher rule quality (see [3,29]).

Recently, Kane et al. [22] relating to more traditional statistical methods advocated that continuous variables (e.g. financial ratios) should be rank-transformed (discretised) to improve their distributional properties in a company failure prediction setting. Within RST based studies on company failure prediction, the CVD process has often been based on expert opinion and also tradition, habits or convention (see [8,12]). How appropriate and consistent the effects of the CVD process from the views of an expert opinion are generally not considered. Articles including [9,39] have identified the need to develop new methods of statistical reasoning (with sparse data). Here, the development is not on the actual CVD techniques employed but a series of measures to describe the effectiveness of any CVD undertaken.

Koczkodaj et al. [24], in a philosophical discussion of RST, considered an information system (set of objects described by characteristics) and propound at what stage does the CVD process effect the objectivity of the information system. Hence, discretisation of data may bring with it subjective uncertainty, with consideration given to the subjective judgements in establishing the boundary points (cut-points) of the defined intervals. This notion is compounded by a motto given in Duntsch and Gediga [15, p. 594], who believes underlying the RST philosophy is "*Let the data speak for themselves.*" It is suggested here that the voice of the data may be muted on the occasion CVD has been applied, with the actual data (continuous values) now described by the intervals within which they exist. This highlights the accuracy versus simplicity problem often described by the Occam (razor) Dilemma (see [9] and references and comments contained therein), whereby here there may be more (accurate) rules on the real data or fewer (simpler) rules using the intervals constructed from CVD. In this paper, while the discretised data may be used knowledge on the positions of the original data in each of the constructed intervals is available and should also be used.

In general, after the utilisation of CVD, the original continuous values may be spread non-uniformly within the different intervals constructed. This spread may involve values near the boundary points of the intervals. Since the continuous values may themselves be estimates (inherently imprecise) then intervals created with a relatively large number of included values near their