A Metric Approach to Supervised Discretization

Dan Simovici*
Richard Butterworth**

*University of Massachusetts Boston
Department of Computer Science, Boston, MA 02125, USA
dsim@cs.umb.edu
**University of Massachusetts Boston
Department of Computer Science, Boston, MA 02125, USA
rickb@cs.umb.edu

Résumé. Nous présentons une nouvelle approche à la discrétisation supervisée des attributs continues qui se sert de l'espace métrique des partitions d'un ensemble fini. Nous discutons deux nouvelles idées fondamentales : une généralisation des techniques de discrétisation de Fayyad-Irani basée sur une distance sur des partitions, dérivée de l'entropie généralisée de Daroczy, et un nouveau critère géométrique pour arrêter l'algorithme de discrétisation. Les arbres de décision résultants sont plus petits, ont moins de feuilles, et montrent des niveaux plus élevés d'exactitude etablis par la validation croisée stratifiée.

1 Introduction

Many machine learning and data mining algorithms can deal only with nominal attributes; however, many data sets of interest have numerical domains and this makes discretization, the conversion from numerical to nominal domains, an important task for data preparation. The literature that deals with discretization is vast and it includes ideas ranging from fixed k-interval discretization [Dougherty *et al.*, 1995], fuzzy discretization (see [Kononenko, 1993]), Shannon-entropy discretization due to Fayyad and Irani presented in [Fayyad, 1991, Fayyad et Irani, 1993], proportional k-interval discretization (see [Yang et Webb, 2003]), or techniques that are capable of dealing with highly dependent attributes (cf. [Robnik et Kononenko, 1995]). The goal of this paper is to introduce a new approach to supervised discretization using the metric space of partitions over finite sets. We present two new basic ideas: a generalization of Fayyad-Irani discretization techniques that relies on a metric on partitions defined by Daróczy's generalized entropy, and a new geometric criterion for halting the discretization process that extends a similar approach proposed by Cerquides and López de Màntaras in [Cerquides et de Màntaras, 1997] using a metric generated by Shannon's entropy.

A *partition* of a non-empty set S is a non-empty collection of non-empty subsets of S, $\pi = \{P_i \mid i \in I\}$ such that $\bigcup\{P_i \mid i \in I\} = S$, and $i,j \in I$, $i \neq j$ implies $P_i \cap P_j = \emptyset$. The set of partitions of S is denoted by $\mathsf{PART}(S)$. For a subset L of M the *trace of the partition* π on the set L is the partition $\pi_L = \{P_i \cap L \mid 1 \leq i \leq k \text{ and } P_i \cap L \neq \emptyset\}$. Daróczy's β-entropy for a partition $\pi = \{P_1, \ldots, P_k\} \in \mathsf{PART}(S)$ is