A Metric Approach to Supervised Discretization

Dan A. Simovici and Richard Butterworth
Univ. of Massachusetts Boston
Dept. of Computer Science
Boston, Massachusetts 02125 USA
{dsim,rickb}@cs.umb.edu

December 15, 2003

Abstract

We introduce a new approach to supervised discretization of continuous-valued attributes that makes use of the metric space of partitions. We present two new basic ideas: a generalization of Fayyad-Irani discretization techniques that relies on a metric on partitions derived from Daróczy's generalized entropy, and a new geometric criterion for halting the discretization process. The resulting decision trees are smaller, have fewer leaves, and display higher levels of accuracy as verified by stratified cross-validation.

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 [DKS95], fuzzy discretization (see [Kon93]), Shannon-entropy discretization due to Fayyad and Irani presented in [Fay91, FI93], proportional k-interval discretization (see [YW03]), or techniques that are capable of dealing with highly dependent attributes (cf. [RK95]). 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 [CdM97] 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 $\mathcal{H}_\beta(\pi) = \frac{1}{2^{1-\beta}-1}\left(\sum_{i=1}^{k}\left(\frac{|P_i|}{|S|}\right)^\beta - 1\right)$, where β is a positive number. It is easy to see that $\lim_{\beta\to 1}\mathcal{H}_\beta(\pi)$ is the Shannon's entropy. The entropy of a partition π_L serves to measure the impurity of the set L relative to the partition π: the larger the entropy, the more L is scattered among the blocks of π. If π, σ are two partitions in $\mathsf{PART}(S)$, the average impurity of the blocks of σ relative to π is the *conditional entropy of π relative to σ*: $\mathcal{H}(\pi|\sigma) = \sum_{j=1}^{m}\frac{|Q_j|}{|S|}\mathcal{H}(\pi_{Q_j})$, where $\sigma = \{Q_1, \ldots, Q_m\}$. López de Màntaras proved that the function $d : \mathsf{PART}(S) \times \mathsf{PART}(S) \longrightarrow \mathbb{R}$ defined by: $d(\pi, \sigma) = \mathcal{H}(\pi|\sigma) + \mathcal{H}(\sigma|\pi)$, where $\mathcal{H}$ is the Shannon entropy is a metric on $\mathsf{PART}(S)$ (see [dM91]).

For $\sigma, \pi \in \mathsf{PART}(S)$, where $\pi = \{P_1, \ldots, P_k\}$ and $\sigma = \{Q_1, \ldots, Q_m\}$, the Daróczy's conditional β-entropy $\mathcal{H}_\beta(\pi|\sigma)$ is given by

$$\mathcal{H}_\beta(\pi|\sigma) = \sum_{j=1}^{m}\left(\frac{|Q_j|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{Q_j}),$$

and thus,

$$\mathcal{H}_\beta(\pi|\sigma) = \frac{1}{(2^{1-\beta}-1)|S|^\beta}\left(\sum_{i=1}^{k}\sum_{j=1}^{m}|P_i \cap Q_j|^\beta - \sum_{j=1}^{m}|Q_j|^\beta\right).$$

A related result obtained in [SJ03] shows that the function $d_\beta : \mathsf{PART}(S) \times \mathsf{PART}(S) \longrightarrow \mathbb{R}$ given by

$$\begin{aligned} d_\beta(\pi, \sigma) &= \mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi) \qquad (1)\\ &= \frac{1}{(2^{1-\beta}-1)|S|^\beta}\left(2 \cdot \sum_{i=1}^{k}\sum_{j=1}^{m}|P_i \cap Q_j|^\beta - \sum_{i=1}^{n}|P_i|^\beta - \sum_{j=1}^{m}|Q_j|^\beta\right). \end{aligned}$$

is a distance; we used it in (Simovici 2003) to obtain small and accurate decision trees.

For $\pi, \sigma \in \mathsf{PART}(S)$ we write $\pi \leq \sigma$ if each block of π is included in a block of σ. If $\pi_1, \pi_2 \in \mathsf{PART}(S)$, then we denote by $\pi_1 \wedge \pi_2$ the partition whose blocks are all non-empty intersections of the form $K \cap H$, where $K \in \pi_1$ and $H \in \pi_2$. The generalized conditional entropy is dually monotonic in its first argument and monotonic in its second, that is $\pi \leq \pi'$ implies $\mathcal{H}_\beta(\pi|\sigma) \geq \mathcal{H}_\beta(\pi'|\sigma)$ and $\sigma \leq \sigma'$ implies $\mathcal{H}_\beta(\pi|\sigma) \leq \mathcal{H}_\beta(\pi|\sigma')$, as we have shown in [SJ03].

If T is a table and A is an attribute of T, we refer to the set of members of the domain of B that occur under B in T as *the active domain of B*; this set is denoted by $\mathsf{adom}_T(B)$, or, if there is no risk of confusion, simply by $\mathsf{adom}(B)$. The partition of the set of tuples of T that corresponds to a partition π of $\mathsf{adom}_T(B)$ is denoted by π_*. A block of π_* consists of all tuples whose B-projections belong to the same block of π.

Discretization of a numeric attribute B involves selecting a set of cutpoints $S = \{t_1, \ldots, t_\ell\}$ in the active domain of the attribute $\mathsf{adom}(B)$, where $t_1 < t_2 < \cdots < t_\ell$. This set of cutpoints creates a partition $\pi^S = \{Q_0, \ldots, Q_\ell\}$ of $\mathsf{adom}(B)$, where $Q_i = \{b \in \mathsf{adom}(B) \mid t_{i-1} \leq b < t_i\}$ for $0 \leq i \leq \ell + 1$, where $t_0 = -\infty$ and $t_{\ell+1} = +\infty$. If the set S consists of a single cutpoint t we shall denote π^S simply by π^t. The discretization process consists of replacing each value that falls in the block Q_i of π^S by i for $0 \leq i \leq \ell$.

Let π_A be a partition of the set of tuples of a table determined by the values of an attribute A. If the list of tuples sorted on the values of an attribute B is $t_1, \ldots, t_n$, define the partition $\pi_{B,A}$ of $\mathsf{adom}(B)$ as consisting of the longest runs of *consecutive* B-components of the tuples in this list that belong to the *same block* K of the partition π_A. The *boundary points* of the partition $\pi_{B,A}$ are the least and the largest elements of each of the blocks of the partition $\pi_{B,A}$. It is clear that $\pi_{B,A*} \leq \pi_A$ for any attribute B.

Fayyad proved that to obtain the least value of the Shannon's conditional entropy $\mathcal{H}(\pi_A|\pi^t_*)$ the cutpoint t must be chosen among the boundary points of the the partition $\pi_{B,A}$, which limits drastically the number of possible cut points and improves the tractability of the discretization [Fay91]. Our main results show that the same choice of cutpoints must be made for a broader class of impurity measures, namely the impurity measures related to generalized conditional entropy. Moreover, when the purity of the partition π^t_* is replaced as a discretization criterion by the minimality of the entropic distance between the partitions π_A and π^t_* (introduced in [SJ03]) the same method for selecting the cutpoint can be applied.

2 A Generalization of Fayyad's Result

We are concerned with supervised discretization, that is, with discretization of attributes that takes into account the classes where the tuples belong. Suppose that the class of tuples is determined by the attribute A and we need to discretize an attribute B. The discretization of B aims to construct a set S of cutpoints of $\mathsf{adom}(B)$ such that the blocks of π^S_* be as pure as possible relative to the partition π_A, that is, the conditional entropy $\mathcal{H}_\beta(\pi_A|\pi^S_*)$ is minimal.

The following theorem generalizes and amplifies Fayyad's result (Theorem 5.4.1 of [Fay91]):

Theorem 2.1 *Let T be a table where the class of the tuples is determined by the attribute A and let $\beta \in (1, 2]$. If S is a set of cutpoints such that the conditional entropy $\mathcal{H}_\beta(\pi_A|\pi^S_*)$ is minimal among the set of cutpoints with the same number of elements, or if $d_\beta(\pi_A, \pi^S_*)$ is minimal among the set of cutpoints with the same number of elements, then S consists of boundary points of the partition $\pi_{B,A}$ of $\mathsf{adom}(B)$.*

To discretize $\mathsf{adom}(B)$ we shall seek a set S of cutpoints such that $d_\beta(\pi_A, \pi^S_*)$ is minimal. Before introducing cutpoints, we have $S = \emptyset$, $\pi^S_* = \omega$, and therefore $\mathcal{H}_\beta(\pi_A|\omega) = \mathcal{H}_\beta(\pi_A)$. When the set S grows the entropy $\mathcal{H}_\beta(\pi_A|\pi^S_*)$ decreases.

Input: A table T, a class attribute A,and a real-valued attribute B.
Output: A discretized attribute B.
Method:
sort table T on attribute B;
compute the set BP of boundary points of partition $\pi_{B,A*}$;
$S = \emptyset$; $d = \infty$;
while BP $\neq \emptyset$ do
    let $t =$ arg min $_{t \in \mathsf{BP}} d_\beta(\pi_A, \pi_*^{S \cup \{t\}})$;
    if $d \geq d_\beta(\pi_A, \pi_*^{S \cup \{t\}})$ then
        begin
            $S = S \cup \{t\}$; BP = BP $- \{t\}$;
            $d = d_\beta(\pi_A, \pi_*^S)$
        end
    else exit while loop;
end while;
for $\pi_*^S = \{Q_0, \ldots, Q_\ell\}$ replace every $q \in Q_i$ by i for $0 \leq i \leq \ell$.

Figure 1: Discretization Algorithm

The use of conditional entropy $\mathcal{H}_\beta(\pi_A|\pi_*^S)$ tends to favor large cutpoint sets for which the partition π_*^S is small in the partial ordered set $(\mathsf{PART}(S), \leq)$. In the extreme case, every point would be a cutpoint, a situation that is clearly unacceptable. Fayyad-Irani technique halts the discretization process using the principle of minimum description. We adopt another technique that has the advantage of being geometrically intuitive and produces very good experimental results.

Using the distance $d_\beta(\pi_A, \pi_*^S) = \mathcal{H}_\beta(\pi_A|\pi_*^S) + \mathcal{H}_\beta(\pi_*^S|\pi_A)$ the decrease in the value of $\mathcal{H}_\beta(\pi_A|\pi_*^S)$ when the set of cutpoints grows is balanced by the increase in $\mathcal{H}_\beta(\pi_*^S|\pi_A)$. Note that initially we have $\mathcal{H}_\beta(\omega|\pi_A) = 0$. The discretization process can thus be halted when the distance $d_\beta(\pi_A, \pi_*^S)$ stops decreasing. Thus, we retain as a set of cutpoints for discretization the set S that determines the closed partition to the class partition π_A. As a result, we obtain good discretizations (as evaluated through the results of various classifiers that use the discretize data) with relatively small cutpoint sets.

3 Discretization Algorithm and Experimental Results

The algorithm is shown in Figure 1. It makes successive passes over the table and, at each pass it adds a new cutpoint chosen among the boundary points of $\pi_{B,A}$. The while loop is running for as long as there exist candidate boundary points and it is possible to find a new cutpoint t such that the distance $d_\beta(\pi_A, \pi_*^{S \cup \{t\}})$ is less than the previous distance $d = d_\beta(\pi_A, \pi_*^S)$. An experiment performed on a syntetic database shows that a substantial amount of time (about 78% of the total time) is spent on decreasing the distance by the

Database	Experimental Results			
	Discretization method	Size	Number of leaves	Accuracy (stratified cross-validation)
heart-c	*standard*	*51*	*30*	*79.20*
	$\beta = 1.5$	20	14	77.36
	$\beta = 1.8$	28	18	77.36
	$\beta = 1.9$	35	22	76.01
	$\beta = 2.0$	54	32	76.01
glass	*standard*	*57*	*30*	*57.28*
	$\beta = 1.5$	32	24	71.02
	$\beta = 1.8$	56	50	77.10
	$\beta = 1.9$	64	58	67.57
	$\beta = 2.0$	92	82	66.35
ionosphere	*standard*	*35*	*18*	*90.88*
	$\beta = 1.5$	15	8	95.44
	$\beta = 1.8$	19	12	88.31
	$\beta = 1.9$	15	10	90.02
	$\beta = 2.0$	15	10	90.02
iris	*standard*	*9*	*5*	*95.33*
	$\beta = 1.5$	7	5	96
	$\beta = 1.8$	7	5	96
	$\beta = 1.9$	7	5	96
	$\beta = 2.0$	7	5	96
diabetes	*standard*	*43*	*22*	*74.08*
	$\beta = 1.8$	5	3	75.78
	$\beta = 1.9$	7	4	75.39
	$\beta = 2.0$	14	10	76.30

Table 1: Experimental Results

last 1%. Therefore, in practice we run a search for a new cutpoint only if $|d - d_\beta(\pi_A, \pi_*^{S\cup\{t\}})| > 0.01d$.

Our discretization algorithm was tested on several machine learning data sets from UCI [BM98] that have numerical attributes. After discretizations performed with several values of β (typically $\beta \in \{1.5, 1.8, 1.9, 2\}$) we built the decision trees on the discretized data sets using the WEKA J48 variant of C4.5 [WF00]. The size, number of leaves and accuracy of the trees are described in below, where trees built using the Fayyad-Irani discretization method of J48 are designated as "standard". The discretization technique had a significant impact of the size and accuracy of the decision trees. The experimental results show that an appropriate choice of β can reduce significantly the size and number of leaves of the decision trees, roughly maintaining the accuracy (measured by stratified 5-fold cross validation) or even increasing the accuracy as shown by

the experiments on the glass data set.

4 Conclusions and Open Problems

With an appropriate choice of the parameter β that defines the metric used in discretization, standard classifiers such as C4.5 or J48 generate smaller decision trees with comparable or better levels of accuracy when applied to data discretized with our technique. We explored only the use of decision trees. Other classification techniques that work with nominal attributes, such as naive Bayes classifiers should also be explored. Also, we intend to examine metric discretization for data with missing values.

References

[BM98] C. L. Blake and C. J. Merz. *UCI Repository of machine learning databases.* University of California, Irvine, Dept. of Information and Computer Sciences, http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998.

[CdM97] J. Cerquides and R. López de Màntaras. Proposal and empirical comparison of a parallelizable distance-based discretization method. In *Proc. of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD'97)*, pages 139–142, Newport Beach,CA, 1997.

[DKS95] J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In *Proc. of the 12th International Conference on Machine Learning*, pages 194–202, 1995.

[dM91] R. López de Màntaras. A distance-based attribute selection measure for decision tree induction. *Machine Learning*, 6:81–92, 1991.

[Fay91] U. M. Fayyad. *On the Induction of Decision Trees for Multiple Concept Learning.* PhD thesis, University of Michigan, 1991.

[FI93] U. M. Fayyad and K. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proc. of the 12th International Joint Conference on Artificial Intelligence*, pages 1022–1027, 1993.

[Kon93] I. Kononenko. Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence*, 7:317–337, 1993.

[RK95] M. Robnik and I. Kononenko. Discretization of continuous attributes using relieff. In *Proc. of ERK-95*, pages 149–152, 1995.

[SJ03] D. Simovici and S. Jaroszewicz. Generalized conditional entropy and decision trees. In *Extraction et Gestion des connaissances - EGC 2003*, pages 363–380, Paris, 2003. Lavoisier.

[WF00] I. H. Witten and E. Frank. *Data Mining – Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, 2000.

[YW03] Y. Yang and G. I. Webb. Weighted proportional k-interval discretization for naive-Bayes classifiers. In *Proc. of the PAKDD*, 2003.