December 2001

Attribute Discretization for Classification

Noel Bryson
Virginia Commonwealth University

Kendall Giles
Virginia Commonwealth University

ATTRIBUTE DISCRETIZATION FOR CLASSIFICATION

Kweku-Muata Bryson
Department of Information Systems
The Information Systems Research Institute,
Virginia Commonwealth University
Kweku.Muata@isy.vcu.edu

Kendall Giles
Department of Information Systems
The Information Systems Research Institute,
Virginia Commonwealth University
kgiles@acm.org

Abstract

Attribute discretization is an important component of the data preparation phase of supervised data mining, and has implications for both the performance of induction learning algorithms and for the use of the resulting decision trees or rule-sets in decision-making. Discretization involves two major decisions: (1) determination of the number of intervals in which the attribute is to be discretized; and (2) the determination of each interval boundary. While most approaches have involved sub-optimal procedures based on different discretization criteria, recently a linear programming approach has been developed that offers optimal solutions for various formulations of the single attribute discretization problem. In this paper we present an exploration of some of the optimal discretizations produced by this method for two data sets that are commonly used in data mining research.

Keywords: Discretization; classification; knowledge discovery; machine learning; linear programming

Introduction

Attribute discretization is one of the important components of the data preparation phase for decision tree (DT) induction and other forms of supervised data mining. Attribute discretization, which involves partitioning the domain of the attribute into a complete set of mutually exclusive intervals, involves two major decisions: (1) the determination of the number of intervals in which the attribute is to be discretized; and (2) the determination of each interval boundary. The attribute discretization step is often necessary for organizational data mining processes because many of the attributes of organizational databases are numeric. Since most decision tree induction algorithms treat each distinct value of an attribute as a discrete value, then using quantitative attributes can lead to explosive growth in the size of the decision tree or rule set and a subsequent degeneration in performance. This effect can be especially problematic for large data sets. Depending on the decisions made regarding the number of intervals and the interval boundaries, the quality of the classification knowledge gained from the discovery process and the accuracy of the results obtained from subsequent application to new events are directly affected.

Importance of Attribute Discretization

Attribute discretization is important for several reasons, including:

1. Discretization has been shown to increase classification accuracy, (Dougherty, Kohavi and Sahami, 1995; Kohavi and Sahami, 1996b; Richeldi and Rossotto, 1995);
2. Discretization reduces the learning effort, the decision tree size, and the number of generated classification rules, thus resulting in a more comprehensible (i.e., simpler) rule set, (Pfahringer, 1995);
3. Attribute (or feature) extraction algorithms which focus on removing redundant and irrelevant attributes make use of the value of the discretization criterion function in determining the attributes that should be included in the minimal set of attributes, (Dash, 1997; Piramuthu, 1999);
4. Discretization simplifies the description of the data, offering meaningful intervals that are derived from relationships in the data rather than being simply manually and subjectively determined;

5. Attribute discretization can itself be considered as knowledge discovery in that critical end-points for the given attribute are exposed, (Kohavi and Sahami, 1996a).

Attribute Discretization Techniques

The attribute discretization problem in supervised learning involves the attempt to partition an attribute's values into a set of mutually exclusive intervals with interval boundaries such that the loss of class/attribute interdependence is minimized. Attribute discretization techniques can be categorized as: error-based (or inconsistency-based) vs. entropy based, global vs. local, dynamic (i.e. during decision tree induction) vs. static (i.e. before decision tree induction), supervised (class information is a factor in discretization) vs. unsupervised (class information is a not a factor in discretization), top-down (splitting intervals) vs. bottom-up (merging intervals). Most previous approaches have involved the use of suboptimal heuristics in the attempt to obtain the optimal partitioning of a given non-class attribute that does not consider other non-class attributes. Initially, simple unsupervised methods were proposed (e.g. equal-width intervals and equal-frequency intervals), but later, various supervised methods were proposed, such as dynamic binarization (Quinlan, 1986), entropy-based methods (Ching, 1995; Stashuk and Naphan, 1992), and Chi-Square-based methods (Kerber, 1992; Liu and Setiono, 1997). These supervised methods integrate the class information of training cases when constructing intervals in order to achieve better classification performance, and are often based on bottom-up, greedy heuristics which could easily miss valuable intervals. Bryson and Joseph (Bryson and Joseph, 2001) proposed a Linear Programming (LP) based approach that could provide optimal solutions for various formulations of the attribute discretization problem including error-based formulations and entropy-based formulations. In this paper we present an exploration of some of the discretizations produced by that method for two data sets that are commonly used in data mining research.

Experimental Experience

Definition of Terms

Let n be the total number of examples in the dataset; $n_{j\bullet}$ be the total number of examples in interval j of the given attribute; $n\bullet s$ be the total number of examples in class s; $n_{j\cap s}$ be the total number of examples in interval j and class s; $p_{j\bullet} = (n_{j\bullet}/n)$ be the estimated probability of being in interval j; $p_{\bullet s} = (n_{\bullet s}/n)$ be the estimated probability of being in class s; $p_{j\cap s} = (n_{j\cap s}/n)$ be the estimated probability of being in interval j and class s; $p_{s|j} = (n_{j\cap s}/n_{j\bullet}) = (p_{j\cap s}/p_{j\bullet})$ be the conditional probability of an example being in class s given that it is in interval j; S be the set of classes. The measure definitions are summarized in Table 1.

Table 1. Measure Definitions

Term	Description
Inconsistency Rate	$IR(g) = \sum_{j\in\Gamma g} \delta_j$ where $\delta_j = (n_{j^\circ} - \text{Max}\{n_{j\cap s}: s \in S\})/n$ and $J_{\Gamma g}$ is the index set of the intervals that are included in the optimal discretization Γ_g that consists of $g = \lvert\Gamma_g\rvert$ intervals.
Information Gain	$IG(g) = -\sum_{s\in S} p_{\bullet s}\log_2(p\bullet s) - \sum_{j\in J\Gamma g} p_{j\bullet}(-\sum_{s\in S} p_{s\lvert j}\log_2(p_{s\lvert j}))$.
Gain Ratio	$GR(g) = IG(g)/SI(g)$, where $SI(g) = \sum_{j\in J\Gamma g} -p_{j\bullet}\log_2(p_{j\bullet})$ is called the *Split Information* for the partition Γ_g with g intervals.
CAMI	$CAMI(g) = \sum_{j\in J\Gamma g}\sum_{s\in S} p_{j\cap s}\log_2(p_{j\cap s}/p_{j\bullet}p_{\bullet s})$.
EffCAMI	$EffCAMI = \text{Max}\{CAMI(g)/SupCAMI(g), g = 2,\ldots, g_{prac}\}$, where SupCAMI(g) = the maximum possible value of CAMI(g).

Software Environment

As part of our research program in decision tree induction we had previously developed software implementations of five entropy-based splitting methods (i.e. Gain Ratio, Conditional Entropy, CAMI, EffCAMI, and CAIR) using the Weka library implementation (www.cs.waikato.ac.nz/~ml/weka) of the well-known C4.5 algorithm, complete with pruning and statistic calculation. The C4.5 algorithm uses Information Gain and Gain Ratio as the decision criteria for choosing an appropriate attribute for branching. In order to test Conditional Entropy, CAMI, CAIR, and EffCAMI, we wrote our own Java programs and classes

to use the C4.5 algorithm structure in the Weka Java library, and substituted the other entropy measures in place of Information Gain and Gain Ratio.

As implemented in C4.5, the decision criterion for selecting an attribute at a node is to pick the attribute with the largest Gain Ratio and whose Information Gain is larger than the average Information Gain of all the candidate attributes at that node. The decision criteria for Conditional Entropy and CAMI are simply to select the attribute with the Min and Max entropy values, respectively. CAIR and EffCAMI however are analogous to the use of Gain Ratio in C4.5. For EffCAMI, the decision criterion is to pick the attribute with the Max EffCAMI value and whose CAMI value is greater than average. The CAIR decision rule is similar to the EffCAMI decision rule. These decision measures are summarized in Table 2.

Table 2. Induction Algorithm Decision Rules for Selecting the Best Attribute

Entropy Measure	Decision Rule
GainRatio	For those attributes whose InfoGain > Average(InfoGain), select the attribute that provides Max(GainRatio).
Conditional Entropy	Select the attribute that provides Min(ConditionalEntropy).
CAMI	Select the attribute that provides Max(CAMI).
CAIR	For those attributes whose CAMI > Average(CAMI), select the attribute that provides Max(CAIR).
EffCAMI	For those attributes whose CAMI > Average(CAMI), select the attribute that provides Max(EffCAMI).

For this study, in addition to the traditional 'dynamic' Gain Ratio approach used in C4.5, we also implemented a 'static' Gain Ratio approach, which treats the discretized continuous attribute as if it were a categorical variable. The 'dynamic' Gain Ratio approach treats the discretized continuous attribute as if it were an ordinal variable. Thus, with the 'static' Gain Ratio approach there would be no further dynamic merging of adjacent intervals by the splitting method, while in the 'dynamic' Gain Ratio approach there could be additional dynamic merging of adjacent intervals.

Description of Test Data

Two data sets were used for demonstrating our LP approach to attribute discretization: the Wisconsin *Breast Cancer* Database (Mangasarian, Setiono and Wolberg, 1990) and the *Iris* data set, both available from the machine learning repository at the University of California at Irvine (Murphy, 1994). The *Breast Cancer* data set consists of 699 examples each belonging to one of two classes, *benign* or *malignant*. Each example is described by nine discrete-valued attributes, in the value domain [1, 10]. A total of 349 examples were selected for the analysis, with enough examples per unique attribute value to ensure that the full range of values was included in our analysis. The *Iris* data set consists of three classes: *setosa*, *versicolor* and *virginica*, with 50 examples from each class. Four numeric, continuous-valued attributes are used to describe each example: *sepal length*, *sepal width*, *petal length*, and *petal width*.

Test Results

For our experiment we did the following: 1) Generated optimal LP-based discretizations for each attribute for different partition sizes; 2) Selected various combinations of partition sizes, although usually the same size was used for all attributes; 3) Applied our modified C4.5 code to generate a DT using various splitting methods (i.e. Static & Dynamic Gain Ratio, CAMI, EffCAMI). Our LP-based discretizations were done for an Inconsistency Rate formulation of the attribute discretization problem (see Table 3 for results) and a CAMI formulation (see Table 4 for results). For both tables the column "# Intervals" indicates the number of intervals for each non-class attribute, while the column "Pre-Discretized" indicates the performance of our DT induction algorithm on the pre-discretized data. Bolded values in Tables 3 and 4 represent instances where discretized approaches produced results as good as or better than the pre-discretized approach.

Inconsistency Rate (I.R.) Discretization

The reader may observe that some combinations of the Inconsistency Rate discretizations give better results than the pre-discretized data for both the IRIS and Breast Cancer datasets. A somewhat surprising result is the relatively good performance

of the Static Gain Ratio method on the Breast Cancer dataset. Given that the Static Gain Ratio method only permits the use of intervals that have been generated in the pre-discretization step, this result indicates that I.R. discretization can result in more valuable intervals that would not be generated by the DT induction algorithm.

Table 3. Inconsistency Rate Discretization

Dataset	# Intervals	Classification Accuracy				
		Gain Ratio			CAMI	EffCAMI
		Pre-Discretized	Static	Dynamic		
IRIS	2223	95.33	**96.00**	**96.00**	**96.00**	**96.00**
	3333		94.00	94.00	94.00	94.00
	4444		94.67	94.67	94.67	94.67
	5555		93.33	**98.00**	**97.33**	**97.33**
	6666		**95.33**	94.67	94.67	94.67
	7777		**96.00**	**96.00**	94.67	94.67
	7733		94.67	94.67	94.00	94.00
Breast Cancer	444444444	72.49	**73.07**	71.06	71.92	71.92
	555555544		71.06	71.92	68.19	69.34
	444333322		**72.49**	71.92	71.35	71.35

CAMI Discretization

The reader may observe that some combinations of the CAMI discretizations give better results than the pre-discretized data for the IRIS dataset but less impressive performances for the Breast Cancer dataset. This result could be partly based on the combinations that we chose for our experiments. The issue of selecting combinations points to one of the problems of the single attribute discretization, where the focus is only on a single attribute. This problem of mixed results in DT induction is not, however, limited to attribute discretization, as previous work on the application of splitting methods in DT induction shows, (Giles, Bryson and Weng, 2001). In the case of splitting methods in DT induction, only the relationship between the given attribute and the target (i.e. class) variable is considered when determining the splits for that attribute.

Table 4. CAMI Discretization

Dataset	# Intervals	Accuracy Rate				
		Gain Ratio			CAMI	EffCAMI
		PreDiscretized	Static	Dynamic		
IRIS	2222	95.33	66.67	66.67	66.67	66.67
	3333		94.00	94.00	94.00	94.00
	4444		**96.67**	**96.00**	**96.00**	**96.00**
	5555		94.67	**96.00**	**96.00**	**96.00**
	6666		**95.33**	**96.00**	**96.00**	**96.00**
	7777		**95.33**	**95.33**	**95.33**	**95.33**
	8888		93.33	96.00	95.33	95.33
Breast Cancer	444444444	72.49	68.77	65.90	68.19	68.19
	555555544		69.63	**72.49**	70.77	71.35

Conclusions

In this paper we have explored the application of two LP-based formulations of the attribute discretization problem on two datasets. The results suggest that this approach can lead to improved classification results, but that it is also important to select good combinations of partition sizes. It should be borne in mind that if the objective is not simply classification accuracy but also to have an interpretable model, the choice of the best combination of partition sizes may be clearer than if the objective is only the best classification accuracy. In the latter case, the fact that the LP-based approach offers optimal partitions for each partition size makes it advantageous over techniques that claim to offer a single, supposedly best discretizations for each attribute. This is because in many cases a combination that involves the 'best' discretization for each attribute based on single attribute discretization may not result in the overall best performance (e.g. the combination "3333" for the IRIS dataset was given as an optimal discretization by the LP formulation, but did not provide optimal classification accuracy).

References

Almuallim, H., and Dietterich, T. "Learning Boolean Concepts in the Presence of Many Irrelevant Attributes," *Artificial Intelligence* (69:1-2), 1994, pp. 279-305.

Bryson, N (K-M)., and Joseph, A. "Optimal Techniques for Attribute Discretizations," *Journal of the Operational Research Society*, in press, 2001.

Ching, J., Wong, A., and Chan, K. "Class-Dependent Discretization for Inductive Learning from Continuous and Mixed-Mode Data," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (17:7), 1995, pp. 631-641.

Dash, M., and Liu, H. "Feature Selection for Classification," *Intelligent Data Analysis* (1), 1997, pp. 131-156.

Dougherty, J., Kohavi, R. and Sahami, M. "Supervised and Unsupervised Discretization of Continuous Features," *Proceedings of the 12th International Conference on Machine Learning*, 1995, pp. 194-202.

Fayyad, U., and Irani, K. "The Attribute Selection Problem in Decision Tree Generation," *Proceedings of the AAAI-92, Ninth International Conference on Artificial Intelligence*, 1992, pp. 104-110.

Giles, K., Bryson, N. (K.-M.), and Weng, Q. "Comparison of Two Families of Entropy-Based Classification Measures with and without Feature Selection," *Proceedings of the Hawaii International Conference on System Sciences (HICSS-34)*, Maui, HI, 2001, pp. 1-10.

Kerber, R. "ChiMerge: Discretization of Numeric Attributes," *Proceedings of the AAAI-92, Ninth International Conference on Artificial Intelligence*, 1992, pp. 123-138.

Kohavi, R., and Sahami, M. "Error-based and Entropy-based Discretization of Continuous Features," *Proceedings of the KDD-96*, 1996a, pp. 114-119.

Kohavi, R., and Sahami, M. "Toward Optimal Feature Selection," *Proceedings of the 13th International Conference on Machine Learning*, 1996b, pp. 284-292.

Liu, H., and Setiono, R. "Feature Selection by Discretization," *IEEE Transactions on Knowledge and Data Engineering* (9:4), 1997, pp. 642-645.

Mangasarian, O.L., Setiono, R., and Wolberg, W.H. "Pattern Recognition Via Linear Programming: Theory and Application to Medical Diagnosis," In *Large-Scale Numerical Optimization*, T. F. Coleman and Y. Li (Ed.), SIAM Publications, Philadelphia, 1990, pp. 22-30.

Murphy, P.M., and Aha, D. W. "UCI Repository of Machine Learning Databases,"), 1994,

Pfahringer, B. "Supervised and Unsupervised Discretization of Continuous Features," *Proceedings of the 12th International Conference on Machine Learning*, 1995, pp. 456-463.

Piramuthu, S. "Feature Selection for Financial Credit-Risk Evaluation Decisions," *INFORMS Journal on Computing* (11:3), 1999, pp. 258-266.

Quinlan, J.R. "Induction of Decision Trees," *Machine Learning* (1), 1986, pp. 81-106.

Richeldi, M., and Rossotto, M. "Class-Driven Statistical Discretization of Continuous Attributes," *Proceedings of the 8th European Conference on Machine Learning (ECML-95)*, 1995, pp. 335-338.

Stashuk, D., and Naphan, R. "Probabilistic Inference Based Classification Applied to Myoelectric Signal Decomposition," *IEEE Transactions on Biomedical Engineering* (39:4), 1992, pp. 346-355.