Zeta: A Global Method for Discretization of Continuous Variables

K. M. Ho and P. D. Scott

Department of Computer Science, University of Essex, Colchester, CO4 3SQ, UK

hokokx@essex.ac.uk and *scotp@essex.ac.uk*

Abstract

Discretization of continuous variables so they may be used in conjunction with machine learning or statistical techniques that require nominal data is an important problem to be solved in developing generally applicable methods for data mining. This paper introduces a new technique for discretization of such variables based on *zeta*, a measure of strength of association between nominal variables developed for this purpose. Following a review of existing techniques for discretization we define zeta, a measure based on minimisation of the error rate when each value of an independent variable must predict a different value of a dependent variable. We then describe both how a continuous variable may be dichotomised by searching for a maximum value of zeta, and how a heuristic extension of this method can partition a continuous variable into more than two categories. A series of experimental evaluations of zeta-discretization, including comparisons with other published methods, show that zeta-discretization runs considerably faster than other techniques without any loss of accuracy. We conclude that zeta-discretization offers considerable advantages over alternative procedures and discuss some of the ways in which it could be enhanced.

1. Introduction

A large number of machine learning and statistical techniques can only be applied to data sets composed entirely of nominal variables. However, a very large proportion of real data sets include continuous variables: that is variables measured at the interval or ratio level. One solution to this problem is to partition numeric variables into a number of sub-ranges and treat each such sub-range as a category. This process of partitioning continuous variables into categories is usually termed *discretization*.

In this paper we describe a new technique for discretization of continuous variables based on *zeta*, a measure of strength of association that we have developed for this purpose. We begin in Section 2 by reviewing existing techniques for discretization and discussing their strengths and limitations. We then proceed in Section 3 to introduce zeta, a new measure of association between nominal variables that is based on minimisation of the error rate when each value of the independent variable must predict a different value of the dependent variable. In the following section we first show how a continuous variable may be dichotomised by a procedure that searches for a maximum value of zeta, and then describe a heuristic extension of this method to partition a continuous variable into more than two categories. In Section 5 we describe a series of experimental evaluations of zeta-discretization including comparisons with other published methods. These results show that zeta-discretization runs considerably faster than other discretization techniques without any loss of accuracy. In the final section we conclude that zeta-discretization offers considerable advantages over alternative procedures and discuss some of the ways in which its could be both improved and extended.

2. Discretization of Variables for Decision Tree Construction

Procedures for constructing classification decision trees using sets of pre-classified examples (Breiman, Friedman, Olshen & Stone 1984; Quinlan 1986) have proved to be among the most effective and useful machine learning techniques. However, such procedures are inherently only applicable to data sets composed entirely of nominal variables. If they are to be applied to continuous[1] variables some means must be found of partitioning the range of values taken by a continuous variable into sub-ranges which can then be treated as discrete categories. Such a partitioning process is frequently termed *discretization*.

A variety of discretization methods have been developed in recent years. Dougherty, Kohavi and Sahami (1995) have provided a valuable systematic review of this work in which discretization techniques are located along two dimensions: *unsupervised* vs. *supervised,* and *global* vs. *local*.

Unsupervised discretization procedures partition a variable using only information about the distribution of values of that variable: in contrast, supervised procedures also use the classification label of each example. Typical unsupervised techniques include equal interval width methods in which the range of values is simply divided into sub-ranges of

[1] Throughout this paper the term ‘continuous’ will be used to refer to variables measured at the interval or ratio level. Ordinal variables form an intermediate case: when the number of distinct values is small, decision trees may be built treating each value as a distinct category; when the number is larger, ordinals should be treated as continuous variables.

equal extent, and equal frequency width methods in which the range is divided into sub-ranges containing equal numbers of examples. More sophisticated unsupervised methods draw on techniques of cluster analysis (Everitt 1980), to identify partitions that maximise within group similarity while minimising between groups similarity (Van der Merckt 1993).

Supervised techniques normally attempt to maximise some measure of the relationship between the partitioned variable and the classification label. Entropy or information gain is often used to measure the strength of the relationship (see for example Quinlan 1986, 1993; Catlett 1991; Fayyad & Irani 1992, 1993). Both ChiMerge (Kerber 1992) and StatDisc (Richeldi & Rossotto 1995) employ procedures similar to agglomerative hierarchical clustering techniques (Everitt 1980): ChiMerge uses χ^2 whereas StatDisc uses Φ (Healey 1990) to determine which groups should be merged. Holte's (1993) 1R attempts to form partitions such that each group contains a large majority of a single classification, subject to a constraint of minimal acceptable group size.

Supervised techniques might reasonably be expected to lead to more accurate classification trees since the partitions they produce are directly related to the class to be predicted. On the other hand one might expect most of the unsupervised techniques to be considerably faster since they involve little more than sorting the data, an operation which is common to all discretization methods.

Global discretization procedures are applied once to the entire data set before the process of building the decision tree begins. Consequently a given variable will be partitioned at the same points whenever it is used in the tree. In contrast, local discretization procedures are applied to the subsets of examples associated with the nodes of the tree during tree construction: consequently the same variable may be discretized many times as the tree is developed and the final tree may include several partitionings of the same variable. The majority of systems using unsupervised methods carry out global discretizations. Examples of supervised global methods include D-2 (Catlett 1991), ChiMerge (Kerber 1992), Holte's (1993) 1R method, and StatDisc (Richeldi & Rossotto 1995). C4.5 (Quinlan 1993,1996) and Fayyad and Irani's (1993) entropy minimisation method both use a supervised technique to perform local discretization. However the majority of supervised techniques could be used for either local or global discretization: for example, Fayyad and Irani's (1993) method has been successfully employed to form global discretizations (Ting 1994; Dougherty et al 1995).

Since local discretization techniques can develop alternative partitionings for different parts of the sample space, one would expect them to be superior to global methods in producing accurate classification trees. However one would also expect to pay a considerable price in execution speed for this improved accuracy since the discretization process may be repeated many times as the tree is built.

Dougherty *et al.* (1995) carried out a comparative study of five discretization procedures using 16 data sets from the UC Irvine Machine Learning Database Repository[2]. The methods compared were two unsupervised global methods (equal width interval procedures), two supervised global methods (1RD (Holte 1993) and Fayyad & Irani's (1993) entropy minimisation)), and C4.5 which is a supervised local method. In all cases the tree was actually constructed by C4.5 but in the four global methods the data was pre-processed using the corresponding procedure to discretize all continuous variables. Somewhat surprisingly Dougherty *et al.* found only small differences, most of which were not statistically significant, between the classification accuracies achieved by resulting decision

[2] Accessible on the World Wide Web at http://www.ics.uci.edu/~mlearn/MLRepository.html

trees. None produced the highest accuracy for all data sets. In particular the local supervised method, C4.5, showed no advantage over other supervised methods: Fayyad & Irani's method achieved the best overall results. Dougherty *et al*. do not report execution times but our own replication (see Section 5) shows that there is a clear trade-off between speed and accuracy: the fastest methods of discretization lead to the least accurate classification trees.

In the rest of this paper we introduce a new procedure for discretization of continuous variables that compares favourably with all the methods discussed in this section. It requires less execution time than other methods and yet leads to classification accuracies that are as high as the best achieved by alternative techniques.

3. Zeta: A New Measure of Association

Our initial attempts to develop a discretization technique that would be both fast and accurate were based upon *lambda,* a widely used measure of strength of association between nominal variables (Healey 1990). Lambda measures the proportionate reduction in prediction error that would be obtained by using one variable to predict another, using a modal value prediction strategy in all cases. Unfortunately lambda is an ineffective measure in those situations were the dependency between two variables is not large enough to produce different modal predictions since in such cases its value is zero.

A closely related measure, which we term *zeta,* has been developed that overcomes this limitation. The fundamental difference between lambda and zeta is that the latter is not based on a modal value prediction strategy: the assumption made in determining zeta is that each value of the independent variable will be used to predict a different value of the dependent variable.

Zeta is most readily understood by first considering the simplest case: using a dichotomous variable A to predict values of another dichotomous variable B. Suppose we have a sample of N items whose value distribution is given in the following 2 by 2 table:

	A_1	A_2
B_1	n_{11}	n_{12}
B_2	n_{21}	n_{22}

where

$$N = \sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}$$

If each value of A is used to predict a different value of B then there are only two possibilities: either $A_1 \rightarrow B_1$ and $A_2 \rightarrow B_2$,or $A_1 \rightarrow B_2$ and $A_2 \rightarrow B_1$. If the former is used then the number of correct prediction would be $n_{11} + n_{22}$; if the latter then $n_{12} + n_{21}$ would be correct. Zeta is defined to be the percentage accuracy that would be achieved if the pairings that lead to greater accuracy were used for prediction. Hence it is defined as follows:

$$Z = \frac{\max(n_{11} + n_{22}, n_{12} + n_{21}) \times 100\%}{N}$$

This definition may be generalised to the case of one k-valued variable A being used to predict the values of another variable B that has at least k values. The more general form of zeta is

$$Z = \frac{\sum_{i=1}^{k} n_{f(i),i}}{N} \times 100\%$$

where f(i) should be understood as follows. In order to make predictions each of the k values of A must be paired with a non-empty set of values of B: these k sets must together form a partition of the set of possible values for B. If B has k distinct values there will be k! ways in which such sets of pairings could be made. One such set of pairing will give the greatest prediction accuracy; call this the *best pairing assignment*. Then $B_{f(i)}$ is the value of B that is paired with A_i in the best pairing assignment.

Note that although the computational effort required to compute Z is extremely modest for small values of k, the computational complexity is O(k!). Hence the measure is only useful for variables with a small number of values. Fortunately this includes a very high percentage of practical cases.

4. Discretization Using Zeta

Having defined a measure of association between two nominal variables we now proceed to describe how this measure may be used to partition a continuous variable. The underlying principle is very simple. In theory, given a k-valued classification variable C, a continuous variable X could be partitioned into k sub-ranges by calculating zeta for each of the possible assignments of the k-1 cut points and selecting that combination of cut points that gives the largest value of zeta. In general such a method will not be practicable because the number of combinations of cut points is extremely large.

4.1. Dichotomising Using Zeta

However it is practicable for the special case of k = 2 since there will only be one cut point. Fortunately this is an extremely common special case in real world classification problems, many of which reduce to a choice between positive and negative diagnoses. If there are N examples in the data set there are at most N - 1 candidate cut points. Zeta can be calculated for every one of these and the point yielding the maximum value selected.

In practice it is not necessary to consider every possible cut point. Fayyad and Irani (1992) have shown that optimal cut points for entropy minimisation must lie between examples of different classes. A similar result can be proved for zeta maximisation. Hence it is only necessary to calculate zeta at points corresponding to transitions between classes, thus reducing the computational cost, particularly when the variables are highly associated.

4.2. Interpreting the Zeta Graph

A graph of zeta values for all candidate cut points illuminates the behaviour of zeta and may give an investigator further insight into the relationship of the two variables concerned. Figure 1 shows a typical zeta graph; this particular example was obtained when dichotomising the Texture variable from the Wisconsin Breast Cancer data set in the UC Irvine Repository.

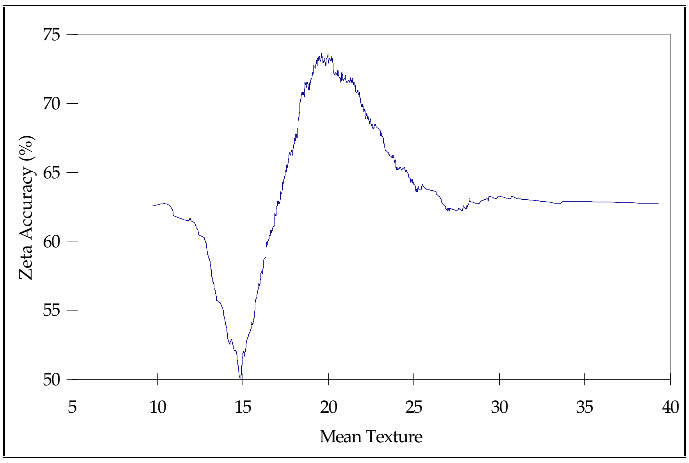

Figure 1: Zeta graph for Texture attribute from the Wisconsin Breast Cancer data in UC Irvine Repository.

Two features of this graph are common to all such zeta graphs:

- A minimum value at the bottom of a deep crevasse. This minimum occurs at the point at which the best pairing assignment changes.
- Both extremities have a zeta value equal to the frequency of the modal class. This is the success rate one would achieve using the optimal strategy to guess the dependent variable without knowledge of the independent variable.

Other features are characteristic of a graph plotted for variables that are strongly related:

- A clearly defined peak rising to a level well above the modal class frequency level. This indicates the optimal position for a dichotomising cut point.
- The graph is fairly smooth. In cases where the dichotomised variable has little predictive value the graph shows numerous reversals of gradient whose magnitude may approach that of the largest peak.

4.3. More Than Two Classes

The dichotomising procedure described above forms the basis of a heuristic method of discretizing a variable into k categories. This is a stepwise hill-climbing procedure that locates and fixes each cut point in turn. It therefore finds a good combination of cut points but offers no guarantee that it is the best. As noted above, examining all possible combinations is likely to be too time consuming.

The procedure for discretizing a variable A into k classes, given a classification variable B which takes k distinct values is as follows. First find the best dichotomy of A using the procedure described above. If k is greater than 2 then at least one of the resulting sub-ranges of A will be associated with 2 or more values of B: use the dichotomising procedure again on such a sub-range to place the second cut point. Proceed in a similar fashion until k - 1 cutpoints have been placed and each sub-range is associated with a different value of B.

This is a heuristic method because once a cut point is placed it is not moved; hence not all cut point combinations are considered. Nevertheless, as some of the results discussed later show, the cut points chosen lead to high predictive accuracy and hence the use of the heuristic is justified.

Attributes	Cut-points	Accuracy (%)
Radius	15.0	89.1
Texture	19.6	73.6
Perimeter	96.3	89.1
Area	693.7	89.4
Smoothness	0.11	67.3
Compactness	0.12	80.0
Concavity	0.09	88.5
Concave points	0.05	91.5
Symmetry	0.21	69.1
Fractal Dimension	0.55	65.2

Table 1: Accuracy obtained using zeta discretization to dichotomise individual attributes in the Wisconsin Diagnostic Breast Cancer data set.

4.4. Computational Complexity

As is the case with almost all discretization procedures for continuous values, the zeta procedure assumes the values have been sorted. This will be on operation of computational complexity O(n log n) where n is the number of examples in the data set. The process of locating each cut point requires a very simple calculation to be performed at most (n - 1) times. Thus if k partitions are formed the operation must be performed at most (k - 1)(n - 1) times. Hence the total computational complexity is O(n log n) + O(kn). When k is small relative to log n, as is usually the case in practice this is approximately O(n log n).

5. Experimental Results

5.1. Discretizing a Single Variable

The first experiments carried out were intended to establish whether the zeta technique partitioned individual variable in a useful fashion. Two types of data were tested: artificial data of known distribution and real data sets containing records of observed phenomena.

5.1.1. Artificial Data Sets

The basic technique employed was to generate data from two or more overlapping normal distributions, assigning each such distribution to a different class. The zeta discretization method was then used to find division points between the classes. The separation of the means were varied. As might be expected, when the separation was relatively large (one standard deviation or more) the cut point was always located very close to the ‘correct’ location, but when the distributions were close together it was positioned less accurately.

5.1.2. Real Data Sets

To investigate the efficacy of zeta in discretizing individual variables in sets of real data, we selected a number of data sets from the UC Irvine repository and investigated how they were divided. As an example, Table 1 shows the results obtained for the continuous attributes in the Wisconsin Diagnostic Breast Cancer data set.

The modal class for this data set forms 62.5% of the total. All the variables have been dichotomised at points that improve prediction accuracy above this baseline. In six cases the accuracy rises to 80% or higher, while the 91.5% accuracy achieved for the single variable ‘concave points’ is approaching the 94.3% achieved by C4.5 (see Table 2) using all the variables.

Thus it can be concluded that the zeta discretization is an effective procedure for locating good cut points within the ranges of continuous variables. We now proceed to consider how it compares with other effective methods.

	C4.5					
Data Set	Continuous	Entropy	1RD	Bin-log l	n-Bins	Zeta
allbp	97.45+-0.10	97.22+-0.16	96.05+-0.32	96.39+-0.32	96.32+-0.13	96.63+-0.23
ann-thyroid	99.61+-0.11	99.38+-0.13	97.64+-0.04	94.06+-0.19	92.72+-0.24	98.38+-0.16
australian	84.93+-0.81	85.65+-1.82	85.22+-1.35	84.06+-0.97	84.93+-0.77	86.38+-0.96
breast	94.28+-0.60	94.42+-0.89	95.13+-0.57	94.85+-1.28	94.85+-0.41	95.85+-0.89
cleve	79.23+-1.63	80.23+-3.25	80.24+-4.15	76.57+-2.60	76.91+-2.11	78.23+-2.37
crx	86.09+-1.11	84.78+-1.94	85.22+-1.93	84.78+-1.82	85.07+-1.80	84.93+-1.99
diabetes	72.66+-1.08	73.70+-0.78	70.45+-1.16	73.44+-1.07	64.85+-1.21	75.13+-1.32
german	71.30+-0.93	72.20+-1.23	70.00+-1.14	72.10+-0.99	71.80+-0.46	73.80+-1.21
glass2	81.00+-2.59	76.67+-1.63	71.23+-5.06	80.42+-3.55	66.86+-2.06	76.14+-1.63
heart	75.19+-1.91	78.52+-1.26	78.52+-0.74	80.74+-1.11	78.52+-1.72	77.41+-3.07
horse-colic	85.87+-1.32	85.60+-1.24	85.60+-1.24	85.33+-1.23	85.60+-1.25	86.15+-1.44
ionosphere	89.45+-1.41	91.15+-1.78	88.88+-1.67	88.60+-1.29	83.17+-2.21	89.72+-2.00
iris	94.00+-1.25	94.00+-1.25	94.00+-1.25	96.00+-1.25	73.33+-2.58	94.00+-1.25
vehicle	73.17+-0.95	68.68+-1.91	66.21+-3.07	68.45+-2.19	62.06+-1.42	69.27+-1.67
waveform-21	76.30+-0.53	74.58+-0.58	52.94+-0.43	70.36+-0.65	74.60+-0.87	76.44+-0.59
Average	84.04	83.79	81.16	83.08	79.44	83.90

Table 2: Classification accuracies and standard deviations using C4.5 (Quinlan 1996) with different discretization methods. Continuous denotes running C4.5 on undiscretized data; Entropy refers to a global variant of Fayyad & Irani’s (1993) method; 1RD is Holte’s (1993) 1R discretizer; Bin-log l and n-Bins use equal width binning; Zeta is the new method proposed in this paper. (c.f. Dougherty et al. 1995).

5.2. Building Decision Trees: A Comparative Study

The next set of experiments was designed to evaluate the performance of zeta in the role for which it was developed: the construction of decision trees. Our experimental procedure was closely modelled on that employed by Dougherty *et al.* (1995) in their comparative study of five discretization techniques.

We compared the five methods considered by Dougherty *et al.* and zeta discretization. C4.5 (Quinlan 1996) was used to construct all of the decision trees. In five of the six cases, the data was first processed by the global discretization procedure and then passed to C4.5. In the sixth case no prior discretization took place; hence the local discretization procedures that form part of C4.5 were used.

The code for zeta discretization was written in C by one of the authors (Ho); the code for C4.5, also written in C, was the version distributed to accompany Quinlan (1993) updated to Release 8 (Quinlan 1996); all the remaining code including both the other four discretization procedures and the code to run the experiments was taken from the MLC++ machine learning library (Kohavi, John, Long, Manley & Pfleger 1994). The data sets used for these experiments were all obtained from the UC Irvine repository. Each set was tested five times with each discretization method.

The results are shown in Table 2. As is to be expected the results for the first five columns are very similar to the results reported by Dougherty *et al.* (1995). The zeta discretization method stands up to the comparison very well. The average accuracy over all the data sets was higher than all the other global methods and only slightly, but not significantly, less than that achieved by C4.5 using local discretization. Thus we can conclude that on average zeta discretization method achieves accuracies at least as good as the best global methods.

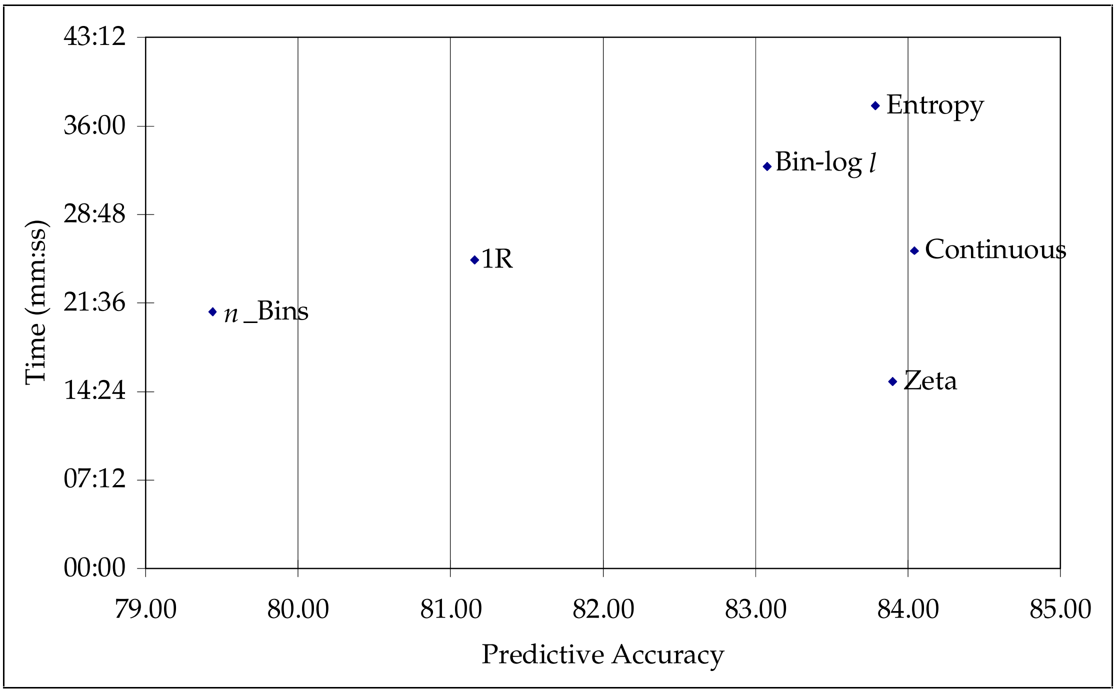

Figure 2: Total execution time for all data sets plotted as a function of average final classification accuracy for different discretization methods (see caption to Table 2).

However, the zeta method is also fast. Figure 2 shows the total execution time required by each of the six methods to complete all the data sets listed in Table 2, plotted as a function of final classification accuracy. It is clear that four of the six data points lie roughly in a straight line, indicating a time accuracy trade-off. Two points lie well below this line: continuous (i.e. C4.5's local method) and zeta. These two methods not only achieve high accuracy but do so in appreciably less time. Thus we also conclude that zeta discretization is the fastest method of achieving high accuracy.

6. Conclusion

These results show that zeta discretization is both an effective and a computationally efficient method of partitioning continuous variables for use in decision trees. Indeed of the methods considered in our comparative study it would appear to be the method of choice.

Although the zeta discretization procedure compares favourably with other methods of partitioning continuous variables so they may be used in constructing decision trees, it is not possible to conclude that it creates the best possible discretization. Indeed there is evidence to suggest that further improvements are possible. For example the zeta procedure achieved an accuracy of 95.85% for the Wisconsin Breast Cancer data set; higher than any other method listed in Table 2. However the documentation accompanying the data cites an accuracy of

97.5% achieved when a hyperplane is constructed to separate the two classes (Wolberg, Street, Heisey & Mangasarian 1995). Thus there is scope for improving zeta discretization.

The technique for partitioning variables into more than two categories is heuristic, since not all combinations of cut points are considered. It is possible that the stepwise method could be improved by using a different strategy for selecting successive cut points. The current system repeatedly creates divisions to that maximise zeta. We have also experimented with a system that repeatedly finds the cut point that best separates the two largest classes within a range or sub-range of values. This produced modest improvements with artificial data but made no difference on real data sets. Another possibility is to investigate whether using zeta as a local rather than global discretization method would lead to any improvements in accuracy.

A more radical enhancement would be to modify the discretization procedure so that it could divide the continuous variable into more categories than there are classifications. This is useful for capturing non-monotonic relationships. For example, suppose one had a data set with a continuous variable indicating age and a dichotomous variable indicating employment status. Young who are still at school and old people who have reached retirement are much less likely to be employed than those whose ages fall in between. Hence the relationship between age and the two employment classes will be better captured by dividing the range into three rather than two sub-ranges. The existing zeta discretization procedure cannot generate more sub-ranges than there are distinct values of the classification variable. We are currently experimenting with a technique that uses secondary peaks on the zeta graph to generate further cut points.

Both zeta graphs and discretization in general have applications beyond the construction of decision trees. The zeta graph shows promise as a useful tool for exploratory data analysis (Tukey 1977). A little experience enables a human investigator to ‘read’ a zeta graph and hence rapidly discover many important aspects of a relationship between two variables. Discretization itself also has an important role in exploratory data analysis: dichotomization is often a useful first step towards discovering the major relationships in a data set (Davis 1971). The zeta procedure will therefore be incorporated into SNOUT, an intelligent assistant for exploratory data analysis currently under development. (Scott, Coxon, Hobbs & Williams 1987).

The results of discretization are often interesting in their own right because category formation is the most fundamental technique that people use to manage the otherwise overwhelming complexity of their experiences. A ‘good’ partitioning is usually one that enables the values of many other attributes to be predicted. All the methods discussed in this paper, including zeta discretization, are concerned with predicting the values of a single classification variable. Thus the most important challenge facing those doing research on discretization procedures is the development of techniques for finding partitions that enable the values of many variables to be predicted.

Acknowledgements

We are grateful to the Economic and Social Research Council’s programme on the Analysis of Large and Complex Datasets for supporting part of the work reported in this paper under grant number H519255030.

References

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984) *Classification and Regression Trees*. Wadsworth, Pacific Grove, CA..

Catlett, J. (1991). On changing continuous attributes into ordered discrete attributes. In *Machine Learning: EWSL-91, Proceedings European Working Session on Learning, Lecture Notes in Artificial Intelligence 482*. pp. 164-178. Springer Verlag.

Davis, J. A. (1971) *Elementary Survey Analysis*. Prentice-Hall Englewood Cliffs, New Jersey.

Dougherty, J., Kohavi, R., & Sahami, M. (1995) Supervised and Unsupervised Discretization of Continuous Features. In *Proc. Twelfth International Conference on Machine Learning*. Morgan Kaufmann, Los Altos, CA.

Everitt, B. (1980) *Cluster Analysis*. Second Edition. Heinemann, London.

Fayyad, U. M., & Irani, K. B. (1992). On the handling of continuous-valued attributes in decision tree generation. *Machine Learning* **8** pp. 87-102.

Fayyad, U. M., & Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *Proc. 13th International Joint Conference on Artificial Intelligence*. pp 1022-1027 Morgan Kaufmann, Los Altos, CA.

Healey, J. (1990) *Statistics: A Tool for Social Research*. Wadsworth, Belmont, CA.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. In *Machine Learning*, **11**, pp. 63-91.

Kerber, R. (1992). ChiMerge: Discretization of numeric attributes. In *Proc. Tenth National Conference on Artificial Intelligence*, pp. 123-128. MIT Press.

Kohavi, R., John, G., Long, R., Manley, D. & Pfleger, K. (1994), MLC++: A machine learning library in C++. In *Tools with Artificial Intelligence*, IEEE Computer Society Press, pp. 740-743.

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

Quinlan, J. R. (1993) *Programs for Machine Learning*. Morgan Kaufmann, Los Altos CA.

Quinlan, J. R. (1996). Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research* **4** pp. 77-90.

Richeldi, M. & Rossotto, M. (1995). Class-Driven statistical discretization of continuous attributes. In *Machine Learning: ECML-95 Proceedings European Conference on Machine Learning, Lecture Notes in Artificial Intelligence 914*. pp 335-338. Springer Verlag.

Scott, P. D., Coxon, A. P. M., Hobbs, M. H. & Williams, R. J. (1997) SNOUT: An Intelligent Assistant for Exploratory Data Analysis. Technical Report CSM-286, Dept. of Computer Science, University of Essex, Colchester, UK.

Ting, K. M. (1994) Discretization of continuous-valued attributes and instance-based learning. Technical Report 491, University of Sydney.

Tukey, J. W. (1977) Exploratory Data Analysis. Addison Wesley, Reading, Mass..

Van de Merckt, T. (1993) Decision trees in numerical attribute spaces. In *Proc. 13th International Joint Conference on Artificial Intelligence*. pp 1016-1021 Morgan Kaufmann, Los Altos, CA.

Wolberg, W. H., Street, w. N., Heisey, D. M. & Mangasarian, O. L. (1995) Computer-derived nuclear features distinguish malignant from benign breast cytology. *Human Pathology* **26** pp 792-796