



The University of Sydney

**Discretization of Continuous-Valued
Attributed and Instance-Based Learning**

Technical Report Number 491

October 1994

Kai Ming Ting

ISBN 0 86758 954 X

**Basser Department of Computer Science
University of Sydney NSW 2006**

Discretization of Continuous-Valued Attributes and Instance-Based Learning

Ting, Kai Ming

Basser Department of Computer Science
University of Sydney, NSW 2006, Australia
E-mail: kaiming@cs.su.oz.au

Technical Report No.491

October, 1994

Abstract

Recent work on discretization of continuous-valued attributes in learning decision trees has produced some positive results. This paper adopts the idea of discretization of continuous-valued attributes and applies it to instance-based learning (Aha, 1990; Aha, Kibler & Albert, 1991).

Our experiments have shown that instance-based learning (IBL) usually performs well in continuous-valued attribute domains and poorly in nominal attribute domains. Cost and Salzberg (1993) have devised the modified value-difference metric (MVDM) that raises the performance of IBL in nominal attribute domains.

This paper explores a way in which continuous-valued attributes and nominal attributes can be treated cohesively in IBL. An algorithm which combines the discretization of continuous-valued attributes and IB1 (Aha, Kibler & Albert, 1991) using the modified value-difference metric is introduced. The empirical results show that the proposed algorithm, IB1-MVDM* achieves a substantial improvement over C4.5 (Quinlan, 1993), IB1 and IB1-MVDM in most of the domains tested. A performance comparison is also made with a naive Bayesian learner (Cestnik, 1990).

Keywords: Instance-based learning, Bayesian learning.

1 Introduction

Virtually no work has been done in treating continuous-valued attributes and nominal attributes in a coherent manner in the framework of instance-based learning. The treatment of continuous-valued attributes in instance-based learning has been acknowledged to achieve satisfactory results. Recent work has been concentrating on improving the treatment of nominal attributes (Cost & Salzberg, 1993) and encouraging results have been attained. As a result of this development, the performance of instance-based learning in nominal attribute domains seems to surpass those in continuous-valued attribute domains. It is the difference in performance in continuous-valued attribute domains and nominal attribute domains that motivates the work reported in this paper.

Relative little research on the discretization of continuous-valued attributes has been reported in the machine learning research literature. Two recent works (Fayyad & Irani, 1993; Van de Merckt, 1993) under the framework of decision trees have benefited the work reported here.

The objective of this paper is to explore a way in which continuous-valued attributes and nominal attributes can be treated cohesively in instance-based learning.

We begin the next section with a brief description of instance-based learning and a recent improvement in the treatment of nominal attributes. Section 3 reports in brief three current methods for discretizing of continuous-valued attributes. A description of the experimental design is given in the next section. Section 5 presents the results and is followed by discussion and future work.

2 Instance-Based Learning

Instance-based learning (IBL) distinguishes itself from other types of learning that induce theories in the forms of decision trees, rules and neural networks. IBL simply stores the training instances and classifies a new instance by predicting that it and its nearest stored instance (according to some distance measures) have the same class (Stanfill & Waltz, 1986; Aha, Kibler & Albert, 1991; Cost & Salzberg, 1993). For a continuous-valued attribute, the value-difference between two values is simply an arithmetic

difference. In a nominal attributes domain, the “overlap” metric is usually used, counting the number of attributes that differ. In order to handle domains that have both types of attributes, the overlap metric is also applied in the same Euclidean distance formulation used for continuous-valued attributes (e.g. the IB series algorithms by Aha, Kibler & Albert, 1991).

Cost and Salzberg (1993) observe that IBL performs relatively poor in nominal attribute domains and identify that it is due to the use of the overlap metric. They use a modified version of the value-difference metric (MVDM), first introduced by Stanfill & Waltz (1986), and empirically show that MVDM is a better metric for nominal attribute domains.

The value-difference distance between two possible values (v_1, v_2) of a specific attribute is defined as follows (Cost & Salzberg, 1993):

$$\delta(v_1, v_2) = \sum_{i=1}^n \left| \frac{C_{1,i}}{C_1} - \frac{C_{2,i}}{C_2} \right|$$

where $C_{1,i}$ is the number of times v_1 was classified into categories i , and C_1 is the total number of times value v_1 occurred.

In the experiment reported in Section 5, two IBL algorithms are used, i.e. IB1 (Aha, Kibler & Albert, 1991) and IB1-MVDM, a modified version of IB1 that incorporates MVDM rather than the overlap metric for nominal attributes.

Our independent experiment (see Table 1) using IB1-MVDM in nominal attribute domains confirms Cost and Salzberg’s claim. Further experiment with continuous-valued domains and comparison with C4.5 seem to reveal two phenomena:

- i. For nominal attribute domains, IB1-MVDM outperforms C4.5 with a large improvement.
- ii. For continuous-valued attribute domains, IB1-MVDM/IB1’s performance is comparable to C4.5.

This observation raises doubt as to whether the metric used thus far for continuous-valued attributes is a good metric. Instead of looking for another metric for continuous-valued attributes, a method is sought to convert continuous-valued attributes to nominal attributes at the outset. We postulate that if a proper method can be found, IB1-MVDM would outperform C4.5 in both continuous-valued attribute and nominal attribute domains.

| <i>Domain</i> | <i>C4.5</i> | <i>IB1-MVDM</i> | <i>IB1</i> |
|------------------|-------------|-----------------|------------|
| bcw | 5.4% | 4.5% | |
| diabetes | 29.3% | 29.8% | |
| waveform | 30.8% | 36.5% | |
| heart | 25.9% | 24.7% | |
| glass | 32.3% | 30.0% | |
| lymphography | 22.1% | 16.4% | 21.7% |
| promoter | 22.6% | 7.1% | 19.3% |
| soybean | 8.1% | 6.0% | 8.5% |
| monks-2 | 35.0% | 25.5% | 29.6% |
| nettalk(phoneme) | 19.7% | 16.2% | 26.1% |

Table 1: Error rates of three algorithms

Another appeal of using IB1-MVDM and converting continuous-valued attributes to nominal attributes at the outset is that it avoids any problems arising from the combination of two different metrics. Because separate metrics are used for continuous-valued and nominal attributes, this may bring about poor results.

Before we show the experiments and their results, let us examine the methods of discretization of continuous-valued attributes in the next section.

3 Discretization of Continuous-Valued Attributes

We describe three discretization methods used in decision trees in this section and use these methods for the experiments described in Section 4.

Fayyad and Irani (1993) devised a method for multi-value splitting for continuous-valued attributes. The same criterion (i.e. information gain) as used in binary splitting (Quinlan, 1986) is employed; however, multi-value splitting is realised by recursively applying the same criterion to the subsets of the first split and so on. A stopping criterion which is based on the minimum description length principle, decides when to stop splitting. This criterion (Fayyad & Irani, 1993) is re-stated as follows:

MDLPC Criterion: The partition induced by a cut point T of attribute A for a set S of N examples is accepted iff

$$Gain(A, T; S) > \frac{\log_2(N - 1)}{N} + \frac{\Delta(A, T; S)}{N}$$

and it is rejected otherwise.

Where

$$\Delta(A, T; S) = \log_2(3^k - 2) - [kEnt(S) - k_1Ent(S_1) - k_2Ent(S_2)]$$

Gain(A,T;S) is the information gain of a cut point T. Ent(S) is the class entropy of a subset S and k is the number of classes in subset S.

Another discretization method is motivated by unsupervised learning. Van de Merckt (1993) makes use of a clustering method to convert continuous-valued attributes into binary splits. Two methods are used; the first is a direct application of clustering method and the second is based on the first criterion but modified with the entropy measure. The two criteria are re-stated as follows:

i. **Unsupervised Monothetic Constrast Criterion:**

$$Contrast(N_1, N_2, A) = \frac{N_1 N_2}{N_1 + N_2} (m_{A1} - m_{A2})^2$$

where N_1, N_2 are number of instances of the resulting binary split, m_{A1} is the mean value for attribute A of N_1 instances.

ii. **Mixed Supervised/Unsupervised Monothetic Criterion:**

$$CE(N_1, N_2, A) = \frac{Contrast(N_1, N_2, A)}{Ent(S)}$$

when Ent(S) is zero, it is replaced by 10E-6.

The desirable split is the cut point that produces the highest Contrast or CE.

We are now ready for the experiment. The experimental design is explained in the following section and the empirical results are shown in Section 5.

4 Experimental Design

The experiment is designed to verify the hypothesis:

The performance of IB1-MVDM can be up-graded by using a proper method of converting continuous-valued attributes to nominal attributes at the outset, and this algorithm compares favourably with IB1 and C4.5 in most domains.

First, a proper conversion method is sought. Six methods have been tried, which include the three methods mentioned in Section 3 and a combination of these methods. Details of the conversion methods are listed below, where in each case, continuous-valued attributes are converted to:

- a. twenty equal interval nominal attributes.
- b. binary attributes according to Unsupervised Monothetic Contrast Criterion (Van de Merckt, 1993).
- c. binary attributes according to Mixed supervised/Unsupervised Monothetic Criterion (Van de Merckt, 1993).
- d. nominal attributes according to MDLPC Criterion, (Fayyad & Irani, 1993)
- e. nominal attributes according to Mixed supervised/Unsupervised Monothetic Criterion with MDLPC Criterion as the stopping criterion.
- f. nominal attributes according to Unsupervised Monothetic Contrast Criterion with MDLPC Criterion as the stopping criterion.

After experimenting with the above six methods, the best method is chosen in combination with IB1-MVDM; and it is designated as IB1-MVDM*. A further experiment is designed to compare IB1-MVDM* with IB1-MVDM, IB1 and C4.5, where all algorithms will be tested without discretization of continuous-valued attributes except IB1-MVDM*.

We are also interested in finding out what are the effects of the discretization on these algorithms. A comparison is made between running with and without discretization for each algorithm. Algorithms run with data conversion are postfixed with “*” to differentiate them from those run without.

5 Empirical Results

The characteristics of the experimental domains and the performance of C4.5 are given in Table 2. These domains include the twelve benchmark domains from Zheng (1993). In addition, two continuous-valued domains (heart disease cleveland and glass), one mixed attribute types domain (automobile) and a nominal attributes domain (nucleotides and codons; see explanation below) are added to the list.

| <i>Domain</i> | <i>%Err C4.5</i> | <i># Ex</i> | <i># Classes</i> | <i># Attr & Type</i> |
|---------------------|------------------|-------------|------------------|--------------------------|
| breast cancer (bcw) | 5.4% | 699 | 2 | 9C |
| diabetes | 29.3% | 768 | 2 | 8C |
| waveform | 30.8% | 300 | 3 | 40C |
| heart disease | 25.9% | 303 | 2 | 13C |
| glass | 32.3% | 214 | 6 | 9C |
| hypothyroid | 0.8% | 3163 | 2 | 18B+7C |
| hepatitis | 23.5% | 155 | 2 | 13B+6C |
| automobile | 24.1% | 205 | 7 | 4B+6N+15C |
| lymphography | 22.1% | 148 | 4 | 9B+9N |
| promoter | 22.6% | 106 | 2 | 57N |
| soybean | 8.1% | 683 | 19 | 16B+19N |
| monks-2 | 35.0% | 169-432 | 2 | 2B+4N |
| nettalk(phoneme) | 19.7% | 5438 | 52 | 7N |
| led7 | 28.6% | 200 | 10 | 7B |
| led24 | 37.9% | 200 | 10 | 24B |
| nucleotides(b) | 37.0% | 20000 | 2 | 60B |
| nucleotides(n) | 35.3% | 20000 | 2 | 15N |
| codons(b) | 25.7% | 20000 | 2 | 320B |
| codons(n) | 36.0% | 20000 | 2 | 5N |

B: Binary, N: Nominal, C: Continuous.

Table 2: Details of experimental domains and performance of C4.5

All these domains are obtained from the UCI Machine Learning Databases except the waveform, led7 and led24 domains, and the last four datasets. These four datasets are actually produced from a single promoter domain, first used by Craven & Shavlik (1993). The original representation is nucleotides(n) which has fifteen 4-value attributes; nucleotides(b) is the binary conversion of the original data. Codons(n)¹ is a transformation from the

¹Three consecutive nucleotides form a codon (Craven & Shavlik, 1993).

original data, which has five 64-value attributes and codons(b) is the binary version.

The first set of eight domains is those with continuous-valued attributes; the first five are all continuous-valued attributes domains and the other three have both types of attributes. The second set constitutes seven all nominal attributes datasets and four binary attributes datasets. The first set of eight domains will be used to verify the hypothesis and finally a complete results on all nineteen datasets will be provided.

All experiments, except in the monks-2, nucleotides and codons datasets, are conducted over 50 runs with randomly selecting 90% of the instances for training and using the remaining 10% as test data. Because a separate set of training data (169) and testing data (432) are given in the monks-2 domain, it is conducted only one run. In the nucleotides and codons datasets, each dataset is split into four equal size and non-overlapping subsets. Four runs are conducted in each of these datasets using each subset (5000) as training data and one of the other subset (5000) as testing data. Only the average results for these runs will be given.

For experiments that employ discretization of continuous-valued attributes in the front-end, they are conducted using the training data first, and then the test data is converted according to the cut points obtained from the training data conversion.

All results shown in the following section are with reference to the error rates of C4.5, tabulated in Table 2, unless otherwise stated.

5.1 Selecting A Discretization Method

Table 3 shows the results of IB1-MVDM* with each of the six discretization methods described in Section 4. The best method in each domain is highlighted. The negative figure indicates that IB1-MVDM* performs better than C4.5 by that percentage and vice versa.

Method (d) achieves four best results and close to the best in the other four domains. Method (f) achieves three best results but performs poorly in the diabetes and glass domains. Method (c) has the best result in the diabetes domain but performs relatively poor in three domains. The other three methods do not produce any best result but have poor performance in at least three domains. Thus, method (d) is chosen to be the method used in combination with IB1-MVDM in the following experiments. These

results also provide a message that there is a discretization method which will make IB1-MVDM* performs better than IB1-MVDM.

| <i>Domain</i> | <i>IB1-MVDM</i> | <i>Discretization Method</i> | | | | | |
|---------------|-----------------|------------------------------|----------|-------------|-------------|----------|--------------|
| | | <i>a</i> | <i>b</i> | <i>c</i> | <i>d</i> | <i>e</i> | <i>f</i> |
| bcw | -0.9 | -1.2 | -1.4 | -0.7 | -0.8 | -1.1 | -1.5 |
| diabetes | +0.5 | +6.5 | -0.2 | -1.4 | -0.7 | +1.9 | +0.9 |
| waveform | +5.7 | -3.2 | -8.2 | -6.3 | -8.2 | -8.1 | -10.1 |
| heart | -1.2 | -4.2 | +3.9 | -3.8 | -6.1 | -4.6 | -3.0 |
| glass | -2.3 | +25.1 | +14.4 | +12.9 | -3.5 | +7.1 | +9.3 |
| hypothyroid | +71.6 | +4.5 | +2.9 | +1.4 | +0.8 | +1.9 | +1.2 |
| hepatitis | -4.4 | -2.6 | -0.6 | -3.5 | -3.2 | -3.2 | -5.4 |
| automobile | +0.6 | +26.6 | +6.0 | +4.4 | -8.5 | -6.3 | -7.2 |

Table 3: Results of six types of conversion and IB1-MVDM

5.2 Performance Comparison

The performance results for IB1-MVDM*, IB1-MVDM and IB1 (with reference to C4.5) are shown in Table 4. The second column indicates that IB1-MVDM* outperforms C4.5 in seven out of eight domains and performs marginally poorer in the hypothyroid domain. Substantial improvements are observed in the waveform, heart and automobile domains. With comparison to IB1-MVDM, IB1-MVDM* is a clear winner in six domains and obtains marginally poorer results in the bcw and hepatitis domains. The same trend is observed in IB1 since IB1 and IB1-MVDM are virtually the same algorithm in the all continuous-valued domains. This experiment supports our hypothesis that IB1-MVDM* is the best algorithm among these four algorithms in most of the domains tested.

The extremely poor performance of IB1-MVDM in the hypothyroid domain needs special attention. An analysis on the MVDM produced for the (original) nine binary attributes indicates that all the value-differences are less than 0.01. When these value-differences are used in conjunction with the continuous-valued attributes in the distance calculation, they have virtually no impact. Thus, the nine binary attributes are deemed irrelevant in IB1-MVDM. Because of the discretization in IB1-MVDM*, the value-differences for the (original) continuous-valued attributes have comparable “weight” to that of the binary attributes. As a result, all eighteen attributes play a part in classification. In fact, in this domain, any of the discretization methods would do better than IB1-MVDM.

| <i>Domain</i> | <i>IB1-MVDM*</i> | <i>IB1-MVDM</i> | <i>IB1</i> | <i>IB1*</i> | <i>C4.5*</i> |
|---------------|------------------|-----------------|------------|-------------|--------------|
| bcw | -0.8 | -0.9 | -0.9 | -2.1 | -0.2 |
| diabetes | -0.7 | +0.5 | +0.5 | -0.5 | -3.5 |
| waveform | -8.2 | +5.7 | +5.7 | -0.2 | +5.5 |
| heart | -6.1 | -1.2 | -1.2 | -4.9 | -2.8 |
| glass | -3.5 | -2.3 | -2.3 | -5.5 | -3.3 |
| hypothyroid | +0.8 | +71.6 | +1.9 | +0.6 | +0.3 |
| hepatitis | -3.2 | -4.4 | -4.2 | -2.9 | -1.2 |
| automobile | -8.5 | +0.6 | +0.1 | -5.3 | +0.9 |

Table 4: Performance comparison

In order to have a clearer picture in comparing the performance difference between the use of converted and original data, Table 5 is produced from the results in Table 4 by subtracting the results of using the converted data from the results of using the original data. Negative figures in Table 5 indicate that the algorithm which uses the converted data is better than that using the original data, and vice versa.

The results clearly show that using the converted data is more desirable for all three algorithms. IB1-MVDM* achieves the greatest improvement, and there is a modest one for IB1*. A mixed performance is observed in C4.5* slightly favouring the converted data with more domains achieving improved results. As a result of this comparison, we will only present results of all algorithms that are combined with discretization method (d) in the next section.

| <i>Domain</i> | <i>IB1-MVDM</i> | <i>IB1</i> | <i>C4.5</i> |
|---------------|-----------------|------------|-------------|
| bcw | +0.1 | -1.2 | -0.2 |
| diabetes | -1.2 | -1.0 | -3.5 |
| waveform | -13.9 | -5.9 | +5.5 |
| heart | -4.9 | -3.7 | -2.8 |
| glass | -1.3 | -3.2 | -3.3 |
| hypothyroid | -70.8 | -1.3 | +0.3 |
| hepatitis | +1.2 | +1.3 | -1.2 |
| automobile | -9.1 | -5.4 | +0.9 |

Table 5: Performance difference between converted and original data for three algorithms

5.3 Overall Results

Table 6 shows the results of all domains for IB1-MVDM*, IB1*, NL* and C4.5*. NL is the naive Bayesian learner that uses the m-estimate (Cestnik, 1990). NL* is a simple classifier but often has good performance; its results are given here for the purpose of comparison. Again, the results shown are with reference to the error rates of C4.5 and the best results are highlighted.

| <i>Domain</i> | <i>IB1-MVDM*</i> | <i>NL*</i> | <i>IB1*</i> | <i>C4.5*</i> |
|------------------|------------------|--------------|-------------|--------------|
| bcw | -0.8 | -2.5 | -2.1 | -0.2 |
| diabetes | -0.7 | -3.8 | -0.5 | -3.5 |
| waveform | -8.2 | -9.6 | -0.2 | +5.5 |
| heart | -6.1 | -7.7 | -4.9 | -2.8 |
| glass | -3.5 | -1.7 | -5.5 | -3.3 |
| hypothyroid | +0.8 | +0.6 | +0.6 | +0.3 |
| hepatitis | -3.2 | -8.1 | -2.9 | -1.2 |
| automobile | -8.5 | +6.6 | -5.3 | +0.9 |
| lymphography | -5.7 | -3.4 | +0.4 | 0.0 |
| promoter | -15.5 | -13.7 | -3.3 | 0.0 |
| soybean | -2.1 | +6.2 | +0.4 | 0.0 |
| monks-2 | -9.5 | +3.2 | -5.4 | 0.0 |
| nettalk(phoneme) | -3.5 | +5.3 | +6.4 | 0.0 |
| led7 | +4.0 | +0.3 | +1.6 | 0.0 |
| led24 | +0.4 | +0.1 | +31.5 | 0.0 |
| nucleotides(b) | -0.3 | -8.1 | -3.0 | 0.0 |
| nucleotides(n) | +2.4 | -6.2 | -1.3 | 0.0 |
| codons(b) | +5.5 | -4.0 | +4.5 | 0.0 |
| codons(n) | -6.4 | -14.4 | -5.8 | 0.0 |

Table 6: Overall results

C4.5* has the best results in the hypothyroid, led7 and led24 domains; but the gain is marginal. IB1-MVDM* achieves the best results in six domains (automobile, lymphography, promoter, soybean, monks-2 and nettalk). IB1* obtains the best result only in the glass domain. NL* attains the best results in six domains or nine datasets (the last four datasets are only different representations of a single domain). NL*'s performance seems to swing at two extremes; in some domains it achieves exceptional good accuracies, e.g. in the hepatitis, nucleotides and codons datasets. Whereas in the automobile, monks-2 and soybean domains, it obtains the worst results. IB1* has moderate performance, i.e. in eleven domains (or fourteen datasets), it has

performance in between the best and the worst. IB1-MVDM* outperforms C4.5 in fourteen out of the nineteen datasets. Considerable big improvements are attained in seven datasets, i.e. the waveform, heart, automobile, lymphography, promoter, monks-2 and codons(n) datasets.

The results also show that IB1-MVDM* doesn't always perform better than IB1* in nominal attributes domains. Examples are the bcw, glass (after discretization) and nucleotides datasets; and IB1-MVDM* seems to have problems in binary domains as well. Looking from the other perspective, IB1-MVDM* seems to deal with irrelevant attributes quite well. This can be seen in the two artificial domains with irrelevant attributes (i.e. led24 and waveform), where substantial gains are achieved when compared with IB1*. Overall, IB1-MVDM* is clearly the best choice among these four algorithms.

6 Discussion and Future Work

The empirical results show that a simple approach in instance-based learning that uses two separate metrics for continuous-valued and nominal attributes can bring about poor performance (an example is in the hypothyroid domain). Converting continuous-valued attributes to nominal attributes at the outset is demonstrated to be a simple yet effective approach. In this way, IBL uses only one metric rather than two separate metrics for continuous-valued and nominal attributes. It provides a coherent approach to the treatment of these two types of attributes.

MVDM seems to be a good metric to deal with irrelevant attributes. Because MVDM establishes that two values (of the same attribute) are similar if they occur with the same relative frequency for all classifications, it provides a means to tolerate irrelevant attributes. Aha (1990) uses an extra parameter, namely weight for each attribute to express the attribute's relevancy; however, MVDM turns out to be a more coherent method which incorporates the capability in the distance metric. The value-difference between two values approaches zero if the attribute is random as they have about the same relative frequency for all classifications.

Cost and Salzberg's claim that the valued-difference metric is superior to the "overlap" metric, is generally true; but there are some cases where the opposite is true. It would be interesting to find out what is the reason and thus possibly provide clues to refine the value-difference metric.

Fayyad and Irani's conversion method emerges as the best in the experiments. However, a clustering method also shows promise. Nominal conversion (f) achieves the best results in three domains (bcw, hepatitis and waveform). This suggests further investigation of unsupervised methods should be carried out.

Though we directly employ Fayyad and Irani's formulation, it is used in a rather different way; we use the formulation only once for the full set of training data, whereas Fayyad and Irani apply it at each level and each branch of the trees.

The experiment that compares the performance difference between converted and original data, indicates that converted data is more desirable for IB1-MVDM, IB1 and C4.5. As Fayyad & Irani (1993) and Van de Merckt (1993) conduct discretization at each level and each branch of the decision trees, it would be interesting to see how these results compare with the discretization done at the outset in the framework of decision trees. Catlett (1991) performed such a comparison by using a discretization method similar to that used by Fayyad & Irani (1993), and found that the combined discretization method and C4 (an earlier version of C4.5) achieve little or no loss of accuracy (and sometimes increased accuracy) over C4 but dramatically increase in efficiency. This findings coincide with the results of our experiment using C4.5.

7 Conclusion

This paper explores a method that combines two recent techniques in machine learning and produces a learning algorithm that is superior to IB1, IB1-MVDM and C4.5 in most of the domains tested. The experiments reported here also identify two areas where IB1-MVDM* could be improved. One area is the refinement of the value-difference metric and the other is the discretization method.

Acknowledgements

This research was partially supported by an Australia Research Council grant (to J.R.Quinlan) and by a research agreement with Digital Equipment Corporation. This author is partially supported by Equity and Merit Schol-

arship Scheme. Numerous discussions with J.R. Quinlan, Mike Cameron Jones, Zijian Zheng, and Alen Varsek have been very helpful. Thanks to J.R. Quinlan and David W. Aha for providing C4.5 and IB1 algorithms. We are also grateful to P.M. Murphy and D.W. Aha for creating and managing the UCI Repository of Machine Learning Databases & Domain Theories

Appendix A

We have attempted to modify the MVDM by using the m-estimate, as used in the naive Bayesian learner (Cestnik, 1990). (the frequency ratios are modified using the m-estimate by setting $m=2$). We repeated the experiment and the relative results of using and not using the m-estimate modification are tabulated in Table A1. A negative figure indicates MVDM with the m-estimate modification is better and vice versa. The results show minor variations with more domains attaining marginal gain than loss.

| <i>Domain</i> | <i>IB1-MVDM</i> | <i>IB1-MVDM*</i> |
|------------------|-----------------|------------------|
| bcw | 0.0 | 0.0 |
| diabetes | 0.0 | +0.2 |
| waveform | 0.0 | -0.9 |
| heart | 0.0 | -0.1 |
| glass | 0.0 | -1.0 |
| hypothyroid | 0.0 | 0.0 |
| hepatitis | -0.3 | 0.0 |
| automobile | -1.7 | -0.4 |
| lymphography | +1.3 | |
| promoter | -0.4 | |
| soybean | 0.0 | |
| monks-2 | 0.0 | |
| nettalk(phoneme) | -0.3 | |
| led7 | 0.0 | |
| led24 | +0.1 | |
| nucleotides(b) | 0.0 | |
| nucleotides(n) | +0.1 | |
| codons(b) | -0.1 | |
| codons(n) | 0.0 | |

Table A1: Results of relative performance of m-estimate
modification in MVDM

Appendix B

Properties of MVDM described by Cost & Salzberg (1992) are:

- a. Two values (of an attribute) are similar if they occur with the same frequency for all classifications.
- b. Mathematical properties:
 1. $\delta(a, b) \geq 0$; for $a \neq b$.
 2. $\delta(a, b) = \delta(b, a)$.
 3. $\delta(a, a) = 0$.
 4. $\delta(a, b) + \delta(b, c) \geq \delta(a, c)$.

There is one important property of MVDM which is not mentioned by Cost & Salzberg (1992), i.e. MVDM is a means to measure attribute relevancy. The following two conditions describe the situations where an attribute becomes irrelevant.

- a. $\{ \delta_p(v_i, v_j) \simeq 0 \mid i \neq j \}$
- b. there exists some k and l with $k \neq l$ such that $\delta_q(v_k, v_l) \gg \max(\{ \delta_p(v_i, v_j) > 0 \mid i \neq j \})$

The first condition states that if all value-differences of an attribute, p , approximate or equal to zero, then this attribute is an irrelevant attribute. The second condition states that if there exists some value-differences of attribute q that are a lot greater than the maximum value-difference of attribute p , then attribute p would become irrelevant.

References

- Aha, D.W. (1990), *A Study of Instance-Based Algorithms for Supervised Learning Tasks*, PhD Thesis, Department of Information and Computer Science, University of California, Irvine, Technical Report 90-42.
- Aha, D.W., D. Kibler & M.K. Albert (1991), Instance-Based Learning Algorithms, *Machine Learning*, 6, pp.37-66.
- Catlett, J. (1991), On Changing Continuous Attributes into Ordered Discrete Attributes, In *European Working Session on Learning*.
- Cestnik, B. (1990), Estimating Probabilities: A Crucial Task in Machine Learning, in *Proceedings of the European Conference on Artificial Intelligence*, pp. 147-149.
- Cost, S & S. Salzberg (1993), A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features, *Machine Learning*, 10, pp.57-78.
- Craven & Shavlik (1993), Learning to Represent Codons: A Challenge Problem for Constructive Induction, *Proceedings of 13th International Joint Conference on Artificial Intelligence*, pp.1319-1324, Morgan Kaufmann.
- Fayyad, U.M. & Irani, K.B. (1993), Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning, in *Proceedings of 13th International Joint Conference on Artificial Intelligence*, pp.1022-1027, Morgan Kaufmann.
- Quinlan, J.R. (1986), Induction of Decision Trees, *Machine Learning*, 1, 81-106.
- Quinlan, J.R. (1993), *C4.5: Program for machine learning*, Morgan Kaufmann.
- Stanfill, C. & Waltz, D. (1986), Toward Memory-Based Reasoning, in *Communications of the ACM*, vol.29, No.12, pp. 1213-1228.
- Van de Merckt, T. (1993), Decision Trees in Numerical Attributes Spaces, in *Proceedings of 13th International Joint Conference on Artificial Intelligence*, pp. 1016-1021, Morgan Kaufmann.
- Zheng, Z. (1993), A Benchmark for Classifier Learning, in *Proceedings of the Sixth Australian Joint Conference on Artificial Intelligence*, pp. 281-286, World Scientific.