Message Length as an Effective Ockham's Razor in Decision Tree Induction

Scott L. Needham
Computer Science and Software Engineering
Monash University
Clayton, Victoria 3168, Australia
sneedham@csse.monash.edu.au

David L. Dowe
Computer Science and Software Engineering
Monash University
Clayton, Victoria 3168, Australia
dld@csse.monash.edu.au

Abstract

The validity of the Ockham's Razor principle is a topic of much debate. A series of empirical investigations have sought to discredit the principle by the application of decision trees to learning tasks using node cardinality as the objective function. As a response to these papers, we suggest that the message length of a hypothesis can be used as an effective interpretation of Ockham's Razor, resulting in positive empirical support for the principle. The theoretical justification for this Bayesian interpretation is also investigated.

"Plurality should not be assumed without necessity" – William of Ockham.

1 INTRODUCTION

Ockham's Razor has long been known as a philosophical paradigm, and in recent times, has become an invaluable tool of the machine learning community. It has been incorporated into many successful machine learning applications, although its validity has remained an area of much debate. As a machine learning heuristic, Ockham's Razor suggests that given a set of equally likely theories about some data, the "simplest" theory is most likely to capture the structure inherent in a problem. Its underlying philosophy has drawn much theoretical support; however, a means for extending this theory to provide sound practical interpretation has proved problematic.

Many statisticians, particularly those of the Bayesian School, have long strived to show that Bayes's theorem represents the mechanism behind Ockham's Razor, and that in fact, it is a consequence of the deeper principles of probability theory. Complementary research has been published supporting this belief, in the form of investigations into the Bayesian (Jefferys and Berger 1991, Good 1968) and classical probabilistic (Forster and Sober 1994) interpretations.

On the experimental front of machine learning, the paradigm has been the target of empirical attack. Murphy and Pazzani (Murphy and Pazzani 1994, Murphy 1995), supported by work from Webb (Webb 1996), have presented a series of papers that attempt to provide empirical evidence against the utility of Ockham's Razor. Experiments in decision tree induction were conducted in which the node cardinality of a decision tree is used as an interpretation of the Ockham's Razor objective function. The relationship between node cardinality and the predictive error of a decision tree was investigated in these papers, apparently putting the Ockham's Razor principle into question.

The current investigation suggests that the node cardinality objective function is a poor, or at least incomplete, interpretation of Ockham's Razor. As an alternative, the inference methods of the Minimum Message Length (MML)[1] principle (Wallace and Boulton 1968, Wallace and Freeman 1987, Wallace and Dowe 1999) provide a practical application of the Bayesian ideals and provides an intuitive interpretation of the Ockham's Razor principle. The MML principle has been successfully applied to a large number of machine learning tasks (Wallace and Dowe 1999, Wallace and Dowe 2000 and their references), which immediately presents a strong argument in favor of Ockham's Razor. The MML message associated with a decision tree is a well studied concept (Wallace and Patrick 1993, Quinlan and R.L. Rivest 1989), and provides a very general interpretation of Ockham's Razor. This paper provides a summary of an extended empirical investigation of the message length interpretation of Ockham's Razor (S.L. Needham, Honours Thesis, CSSE, Monash University, 2000).

[1] The similar, but independent methods of Minimum Description Length (MDL) inference (Rissanen 1978) would give a similar interpretation of Ockham's Razor.

2 PREVIOUS EXPERIMENTAL EVIDENCE

In this section, the characteristic experiment investigated in the work of Murphy and Pazzani (Murphy and Pazzani 1994) is replicated. The hypothesis space of binary decision trees was used to learn the binary logic concept $(XYZ)|(AB)$ without noise and without dummy attributes. The experiments involved 100 trials being run, each creating a training set by randomly choosing without replacement 20 of 32 ($= 2^5$) possible training examples. The remaining 12 examples were used as a test set. For each trial, every consistent decision tree (those with only pure leaves showing all things in the same class) was created, and the average error rate made by trees for each node cardinality was computed. Figure 1 plots the mean and 95% confidence interval of the average "right"/"wrong" errors as a function of the node cardinality. The average number of trees found to have each node cardinality is also plotted.

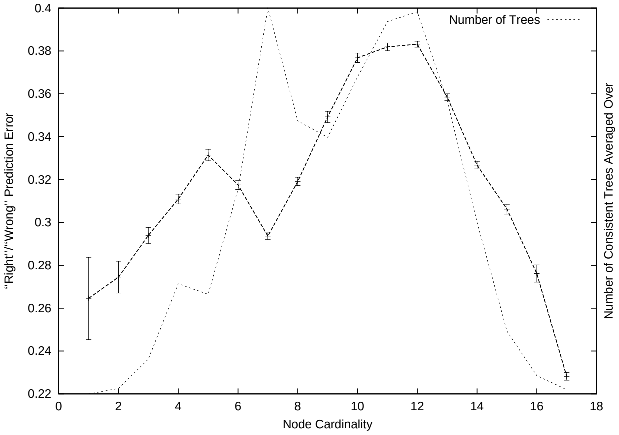

Figure 1: Node cardinality vs. Prediction Error

The results in Figure 1 compare closely to those found by Murphy and Pazzani (Murphy and Pazzani 1994), and indicate that the node cardinality objective function does not provide positive support for Ockham's Razor. Figure 1 suggests that on average, trees with node cardinality 7 have a lower "right"/"wrong" error rate on unseen data than trees with lower node cardinalities. If we accept Murphy and Pazzani's interpretation of Ockham's Razor, this evidence suggests a violation of Ockham's Razor.

3 THEORETICAL INTERPRETATION

The work of Murphy and Pazzani (Murphy and Pazzani 1994) suggests a practical interpretation of Ockham's Razor which, on the surface, does not seem unreasonable. The results described above agree with those of Murphy and Pazzani, although it is our belief that the node cardinality of a decision tree is a poor interpretation of Ockham's Razor. This section takes on a theoretical investigation of the paradigm, with the intent of finding a more appropriate Ockham's Razor objective function.

3.1 A BAYESIAN INTERPRETATION OF OCKHAM'S RAZOR

Bayesian philosophy requires that hypotheses have associated prior probabilities, which is the essence of its approach to statistics. Good tells us that"'Ockham's Razor' states that if two hypotheses H and H_1 explain the facts equally, meaning $P(E|H) = P(E|H_1)$, then the simpler of the two is to be preferred", (Good 1968). We can see from Bayes's theorem, $P(H|D) = Pr(D\&H)/Pr(D) = Pr(H)Pr(D|H)/Pr(D)$, that this preference is equivalent to the choice of the more probable hypothesis. The Minimum Message Length principle presents a Bayesian method, which uses subjective priors to make this choice.

In the general machine learning problem, we are given a set of data D, from which we wish to infer a hypothesis, H. When looking for the most appropriate hypothesis for some given data, Bayes's theorem suggests that we choose the hypothesis with the highest posterior probability, $P(H|D)$, or equivalently, that theory which maximizes the product of the prior probability of the theory, $P(H)$, with the probability of the data occurring in light of the theory, $P(D|H)$. In terms of Ockham's Razor, a good theory for some data will have an accordingly high prior probability and a good likelihood "fit".

The MML principle provides a theoretical and somewhat intuitive means for making the connection between Ockham's Razor and a corresponding quantitative metric. As above, we can regard the problem of maximizing the posterior probability, $Pr(H|D)$, as one of choosing H so as to maximize $Pr(H).Pr(D|H)$. Since $-\log_2(Pr(H).Pr(D|H)) = -\log_2(Pr(H)) - \log_2(Pr(D|H))$, maximizing the posterior probability, $Pr(H|D)$, is equivalent to minimizing

$$MessLen = -\log_2(Pr(H)) - \log_2(Pr(D|H)),$$

the length of a two-part message conveying the theory, H, and the data, D, in light of the theory. Hence the name "minimum message length" (principle) (Wallace and Boulton 1968, Wallace and Freeman 1987, Wallace and Dowe 1999) for choosing a theory, H, to fit observed data, D. The part of the MML message expressing the hypothesis can be obtained by creating a Shannon optimal code for the language describing the set of hypotheses and then constructing the message from this code (Wallace and Freeman 1987).

As stated at the start of this section, many regard Ockham's Razor to be primarily concerned with hypotheses that have equal likelihood given some data (Good 1968). In these cases, the MML principle suggests that the hypothesis with the shortest encoding is most likely to be the best predictor of future data.

3.2 ACQUISITION OF PRIORS FOR BAYESIAN INFERENCE

Referencing Bayes's Theorem as it applies to the inference of hypotheses, finding the posterior probability of a hypothesis given some data requires the probability of that hypothesis a priori. The prior probability of a hypothesis is usually interpreted as the probability that the hypothesis describes the true source of a particular data set. It is clear that this probability distribution over all hypotheses is very difficult to calculate. Even in restricted hypothesis spaces, the task is usually intractable and approximate prior probability distributions are used. As discussed in Section 3.1, the MML techniques utilize a Shannon optimal code for a given hypothesis space, using it to construct an encoding for each hypothesis. The message length for each hypothesis serves as an approximation to the negative logarithm to the base 2 of the hypothesis' true prior probability.

A common argument against Bayesian inference methods revolves around the selection of ludicrous prior probability distributions. For example (similar to that given by Domingos (Domingos 1999)), suppose we gave one particular decision tree with one million nodes a prior of 0.5, and then allocated equal prior probability to all remaining trees. This would result in the MML inference techniques and most Bayesian inference techniques often inferring this hypothesis given a variety of data. It has been argued (Domingos 1999), that by having a decision tree of such a large node cardinality being selected, that Ockham's Razor has been violated. Bayes's theorem, in its simplest form, makes no restriction in principle on the type of prior probability distribution that is chosen. However, Bayesian philosophy suggests that the selection of the prior probability distribution is important. The selection of a prior probability distribution as described above would only ever be made if we truly believed that this hypothesis, with one million nodes, actually did occur with probability of 0.5. In this case, in a message length framework, we would describe the hypothesis with an optimal encoding of one bit. This does not disagree with Ockham's Razor in any way, as the most probable hypothesis has the simplest description. Assigning ludicrous prior probabilities to hypotheses, disregarding our belief in their true prior probabilities, would have to be very strongly questioned in practice. Such attempted sabotage contradicts both Bayesian philosophy and basic intuition. The use of misrepresentative priors in no way undermines the effectiveness of Bayesian inference, which endeavors to use plausible rather than ludicrous priors. See (Lindley 1972, Bernardo and Smith 1994, Solomonoff 1999, Wallace and Dowe 1999) for some of the very broad discussion on the selection of Bayesian priors and Section 4.1 for an analysis involving more plausible priors.

4 PRACTICAL APPLICATION OF OCKHAM'S RAZOR

In the previous section, the groundings of Ockham's Razor in the theoretical field of Bayesian statistics was discussed. The practical validation of these ideas is now investigated on the restricted search space of decision trees.

4.1 DECISION TREE ENCODING

For decision trees, there are four elements to consider when encoding their message in an MML framework (Wallace and Patrick 1993, Quinlan and Rivest 1989). These are:

$1a$: the encoding of the structure of the tree - this involves encoding whether a node is a leaf or an internal node.

$1b$: is the labeling of each internal split node with an attribute.

$1c$: in each leaf node there is the encoding of the probabilistic prediction associated with each category. This completes the encoding of the hypothesis. [2]

2 : Finally, there is the encoding of the category of each thing, using for each a code based on the probabilistic prediction associated with the thing's true category.

A complete investigation of this encoding is described in Wallace and Patrick (Wallace and Patrick 1993).

4.2 MESSAGE LENGTH AS AN EFFECTIVE OCKHAM'S RAZOR

We now re-visit the investigation taken on by Murphy and Pazzani (Murphy and Pazzani 1994), substituting the node cardinality objective function with that of the message length measure. Figure 2 displays the relationship found between the complete message length (parts 1a,1b,1c and 2) and "right"/"wrong" percent-

[2] The order in which these components are arranged in the message need not necessarily be 1a, 1b, 1c.

age error. The distribution of trees over the message length domain is also plotted. The message length objective function has continuous values, and for this reason the error is averaged over a number of intervals of message length. Twenty equal intervals have been used for the purpose of good visual comparison to the node cardinality results in Figure 1, where the maximum cardinality was 20. A notable shift in the distribution of trees from Figure 1 is seen when the message length is applied in Figure 2. The correlation between message length and percentage error does not follow a smooth monotonic curve, although it certainly shows a positive correlation as indicated by our least-squares regression fit.

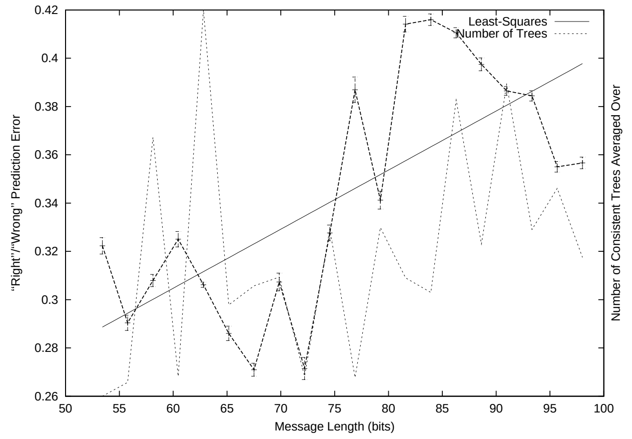

Figure 2: Message Length vs. Prediction Error

The experiments conducted by Murphy and Pazzani (Murphy and Pazzani 1994) consider only consistent decision trees, for which we presume the following justification. Re-iterating Good's definition of Ockham's Razor, given "two hypotheses H and H_1 explain the facts equally, meaning $P(E|H) = P(E|H_1)$, then the simpler of the two is to be preferred", (Good 1968). Taking this interpretation, the set of consistent decision trees is certainly a set for which Ockham's Razor applies, with $P(E|H_i) = P(E|H_j)$ for all H_i and H_j in the set, since $P(E|H) = 1$ for all H in the set.

Bayes's theorem suggests that if we have a set of hypotheses with constant likelihood, then the posterior probability of a hypothesis, given some data, becomes a simple multiple of its prior probability. In terms of the message length of a hypothesis, this translates to selecting the hypothesis with the shortest encoding or, in other words, the hypothesis with the shortest "first part" (1a,1b and 1c) of the MML message. In practice however, this is not precisely the case. The definition of a consistent decision tree requires that it have only pure leaves, that is, the tree makes predictions over the data with probabilities 1 and 0. In a simple example, suppose we use one such consistent tree to construct a Huffman code for the purpose of transmitting future data. In this case, any incorrectly classified data would require an infinite number of bits to be transmitted. Clearly, 100% pure predictions should be made with extreme care, if at all.

In practice, MML techniques make predictions on future data with some probability greater than zero. If we have n_m training data for class m then we predict class m with probability, $p_m = (n_m + \frac{1}{2})/(N + M/2)$, where N and M are the number of training examples and classes respectively (Wallace and Freeman 1987, Wallace and Dowe 2000). It can be seen, that the only time 100% pure predictions would be made is when an infinite and pure training data set is available. As a result, it is found that part 2 of the MML message, the encoding of the data given the hypothesis, is not constant across the space of consistent decision trees, but is a function of the distribution of data in the leaves. However, the contribution of part 2 on the complete MML message is small and near constant.

The idea of judging a prediction based on its encoding cost can be extended to the context of the current problem. It can be strongly argued that the logarithm of probability bit score provides a better discriminator of the performance of a hypothesis. In this case, instead of using a percentage error, we score each hypotheses by giving it $-\log_2(p)$ bits for each test data item, where p is the probability with which the hypothesis predicted the actual class of the data item - see (Dowe et al. 1996) and its reference list for a discussion. The nature of this measure suggests that hypotheses with small bit cost are good predictors of the data, so again strong support for Ockham's Razor would be indicated by a smooth monotonically increasing plot. Figure 3 presents the previous results implementing the logarithm of probability bit score in place of the "right"/"wrong" predictive error. The results found appear to be very similar to those found with the "right"/"wrong" predictive error. When using consistent trees, the leaf distributions are near Bernoulli with $p = 1$, with this simple distribution the range of bit scores is also simple usually only taking on values near 0 and 1.

The experimental results for the message length objective function in Figures 2 and 3 seem to give some positive support for Ockham's Razor, with a positive correlation being seen in each case. However, there are some characteristics of the results appearing consistently, which do not allow for any conclusive claims to be made about the validity of Ockham's Razor. For example, the results in Figures 2 and 3 display that at some points along the message length axis, the average performance (i.e. both "right"/"wrong" predictive accuracy and the logarithm of probability of bit score)

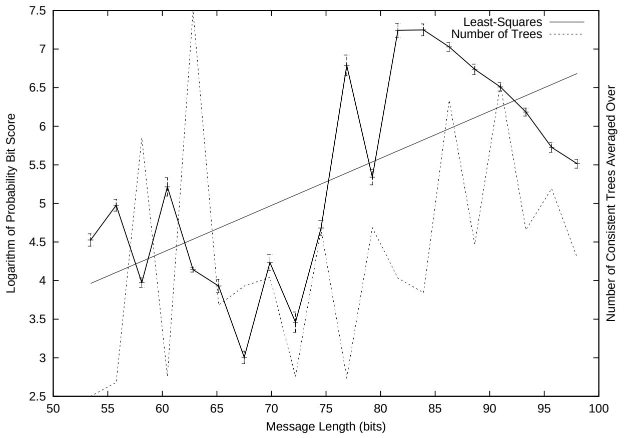

Figure 3: Message Length vs. Bit Score

drops to a value below that found for the shorter message lengths. These results are open for the criticism that Murphy and Pazzani discuss in their investigation of the node cardinality objective function. This problem is most prominent for the average performance results for message lengths of greater than 80 bits, where a continuous decrease in average performance is seen for over five intervals of message length. It is believed that these results many be dependent of the experimental conditions, for which an indepth investigation is made in the following section.

5 DISCUSSION OF RESULTS AND LEARNING TASK

The message length objective function has been shown to provide better support for Ockham's Razor than the previously suggested node cardinality objective function (Murphy and Pazzani 1994). However, this improvement has not been sufficient to provide undisputed evidence for the validity of our interpretation of Ockham's Razor. It could be argued that even though the message length objective function was unable to provide clear support for Ockham's Razor in these experiments, that in fact no objective function will perform well on the proposed learning task. As an alternative example to make this point clear, suppose that we applied our objective function to the task of inferring a hypothesis about some large sample of random noise. Of course the results will provide no support for Ockham's Razor, but by no means could we argue that this is evidence against Ockham's Razor. The suspicion that these poor results could be related to the experimental conditions is investigated in this section.

In the experiments conducted involving message length, it was found that for the trees of large message length, the results were not in favor of Ockham's Razor. The relationship between message length and the performance measures displayed a negative gradient, suggesting that on average, trees of relatively longer message length were better predictors of the future data. This trend was seen consistently in the results and seems to be a consequence of the small training sets. Noting that the MML approximation of the probability associated with a class is never zero, the class probabilities for a binary leaf with one training data thing are 3/4 and 1/4. That is, the MML approximation of probability suggests that even though there is no data supporting a class, there is insufficient data to make a pure prediction. This not only reduces the penalty for incorrectly classifying a test example, but also reduces the encoding cost of the leaf distribution. As a result it is found that large trees, with many leaves containing one data example, are found to make relaxed predictions and thus incur relaxed penalties for incorrect classifications. This is not a problem with the MML approximations as the choice not to make pure predictions with one data example appears reasonable. The problem appears to be related to the insufficient sample sizes. Methods for creating larger training data sets are investigated in the next section.

A second concern with the investigation was with the choice to only investigate consistent decision trees. This decision was made for the purpose of investigating Good's interpretation of Ockham's Razor and to provide a comparative investigation with the work of Murphy and Pazzani (Murphy and Pazzani 1994). However, it appears that this restriction is not necessary. The argument made by Murphy and Pazzani for using consistent decision trees is that typically (Quinlan 1986, etc.) decision tree induction methods use consistency as a stopping criterion. For many methods, this is the case, although many induction techniques (Wallace and Patrick 1993, Quinlan and Rivest 1989, Uther and Veloso 2000) do not restrict their search space to consistent trees. This leads to the extended investigation, involving the complete space of decision trees, taken on in the following section.

5.1 OCKHAM'S RAZOR; AN ALTERNATIVE INTERPRETATION

Ockham's Razor has been a debated topic for centuries, with the debate extending to the disagreement on the words that Ockham actually spoke. Clearly, this makes constructing an interpretation of Ockham's Razor in the context of machine learning a difficult task. Referring to its commonly accepted translation: "plurality should not be assumed without necessity", we find that a clear mathematical interpretation is not obvious. Ockham's Razor seems to suggest that we should prefer a simpler hypothesis while the benefit of the reduced complexity is not outweighed by a decrease in the goodness of "fit" of the hypothesis. That is, we prefer a simpler hypothesis while the combined complexity of its description and the data given it, is shorter than that of the current hypothesis. This new interpretation of Ockham's Razor is a generalization of that given by Good (Good 1968), the interpretation is equivalent to Good's in the case where the set of hypotheses considered has constant likelihood. The Minimum Message Length principle accesses this trade off between the complexity of the hypothesis and the likelihood of the hypothesis given some data. This is achieved by comparing hypotheses using the two-part encoding of the hypotheses and the data given the hypothesis (refer Section 3). In this section, the performance of this new interpretation will be investigated through a series of experiments.

The major consequence of this new interpretation of Ockham's Razor on the experimental investigation is that we are now concerned with the complete space of decision trees, and not only those consistent with the test data. Experimentation with the node cardinality objective function requires a set of decision trees that have a constant likelihood given some data (e.g. the set of consistent trees). This is because the node cardinality objective function does not incorporate a measure of the goodness of "fit" of a hypothesis. As a result it will make no differentiation between two trees with equal node cardinality even if one correctly classifies all of the test data and the other does not correctly classify a single example.

The practical investigation of this new interpretation of Ockham's Razor follows a similar path to that taken in the previous section. Figure 4 displays the results for experiments incorporating the complete space of decision trees. The experiments are otherwise identical to those conducted in Section 4.2, with the data having no noise or dummy variables. The results found demonstrate a smoother relationship between the message length and the performance measures. This is most likely the result of the greatly increased number of experimental points used to create the plots, as a huge number of trees that are not consistent with the data are now included in the averaged results. Nevertheless, these results do not provide clear support for Ockham's Razor. A clear drop in the predictive error and bit score is seen for trees with message lengths of around 65 bits, similar to that seen in the previous experiments. Also, a trend of decreasing average predictive error and bit score starting for trees with message length of around 80 bits and continuing to the trees of maximum message length is again seen. In the first case we can offer little explanation for this evidence against Ockham's Razor. In the second case, as discussed in the previous section, we believe the trend to be the result of the small size of the data set.

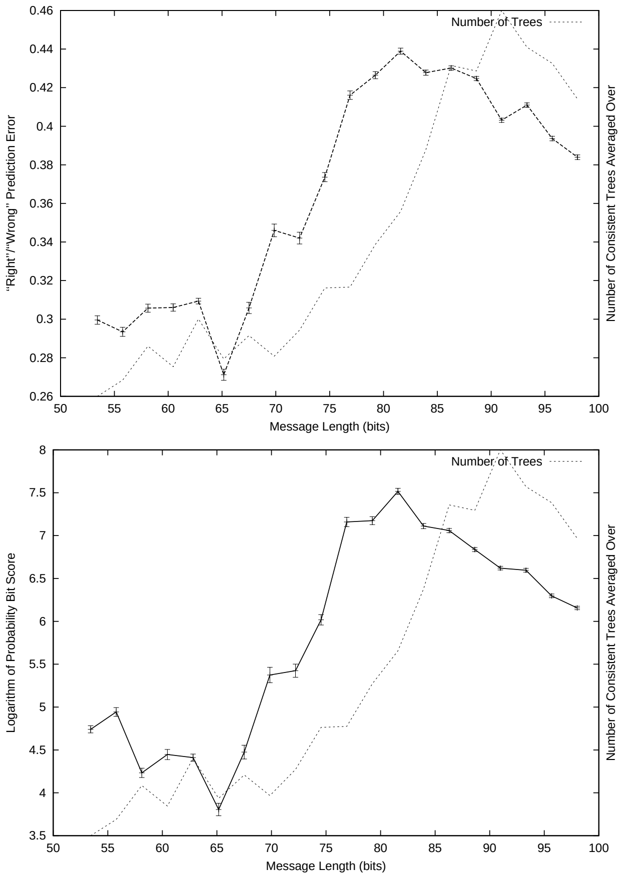

Figure 4: Results With Complete Search Space

5.2 ALTERNATIVE TRAINING DATA GENERATION ALGORITHM

As discussed in Section 5, it is felt that the size of the training data set is not adequate for the leaf distributions to make accurate predictions about the data. In the previous section the complete space of decision trees was introduced into the investigation. This section extends these experiments by incorporating a new training data generation algorithm.

It should be possible to use some arbitrary number of training examples for the purpose of decision tree inference. When training data is taken from a "real world" data source, in many cases, the number of samples that can be attained is only constrained by the time that is spend gathering the data. In contrast to the current method for attaining training data, in "real world" data samples it is expected that the data examples may occur many times and with different frequency. Also, the data is affected by the measurement error found in the experimental equipment and often the attributes relating to the data are not obvious. Using this “real world” model, a new method for data generation is suggested where by an arbitrary number of training data examples can be created. The method is given by the simple algorithm:

```
repeat until sufficient examples are created {
1. Randomly select a permutation of the attributes to create a data thing.
2. Evaluate the class of the data thing and with some probability assign a noisy class to the data thing.
3. Add the data thing to the training set.
}
```

When a data thing is affected by noise, the class associated with its attribute vector is assigned randomly without reference to the true value. This means that as the probability of noise approaches 1 the data becomes completely random. A “dummy” variable is incorporated into the data by simply included it in the attribute vector of each data thing and assigning it a random value, thus it provides no information about the true class of the data thing. In a “real world” learning task, the performance of the inferred hypotheses is assessed by testing the hypotheses on future data samples. When using an artificial learning task, the hypotheses can be evaluated using the complete uncorrupted data set. With this algorithm we do not explicitly withhold a subset of the data set for testing, although there is some probability proportional to the size of the data set, that an example will be excluded.

Figure 5 displays the results found when a training set of 200 examples is used with 0.3 probability of noise and 1 dummy variable . In Section 2, it was discussed that the ideal support for Ockham’s Razor would consist of a smooth monotonically increasing curve over the entire message length domain. This is clearly demonstrated in Figure 5, which satisfies these requirement to near perfection, in contrast to the previous results shown in Figure 4. Results of near this quality are seen for data sets as small as 30 training examples and for noise levels as high as 0.5 probability.

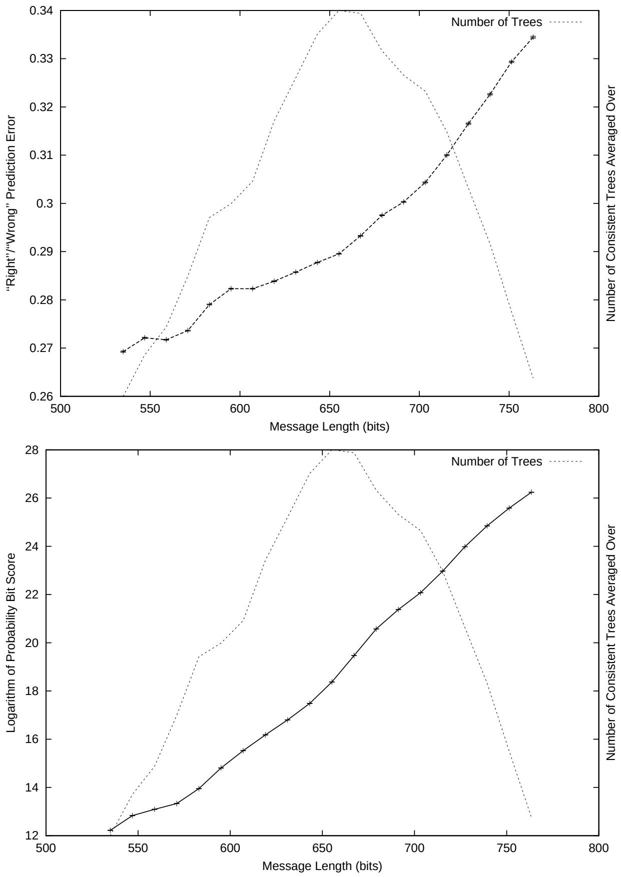

Figure 5: Data With 0.3 Probability of Noise and 1 Dummy Variable

6 CONCLUSION

The focus of this paper has been primarily to provide a response to the growing number of empirical investigations (Murphy and Pazzani 1994,Murphy 1995,Webb 1996,Domingos 1999) that have appeared to discredit the Ockham’s Razor principle as a machine learning objective function. In particular, we have focused on the investigation taken on by Murphy and Pazzani (Murphy and Pazzani 1994), who in their work propose node cardinality as an Ockham’s Razor objective function for the induction of decision trees. We have proposed that the node cardinality objective function is a incomplete interpretation of Ockham’s Razor, and that instead the MML message length is an effective alternative in decision tree induction.

From the experiments conducted using the learning task proposed by Murphy and Pazzani (Murphy and Pazzani 1994), it was shown that the message length interpretation of Ockham’s Razor clearly outperformed that of the node cardinality. The results, however still did not provide undisputed evidence for the Ockham’s Razor principle. We proposed a new interpretation, which is closely related to the MML and Bayesian interpretations of Ockham’s Razor: we prefer a simpler hypothesis while the combined complexity of its description and the data given it, is shorter than that of the current hypothesis. With this interpretation, the complete space of decision trees was included in the investigation. Nevertheless, while the experimental investigation of this new interpretation yielded improved support for Ockham’s Razor, the evidence was still inconclusive. A new means for creating training data is proposed that is based on a “real world" model, which includes repeated data points, noise and dummy variables. Experimentation with this new data generation algorithm produced exceptional results with respect to the "right"/"wrong" prediction error and logarithm of probability bit score.

In future work, it is hoped that the strong support shown for Ockham's Razor in the current investigation can be extended to the investigation of other learning tasks, and alternative hypothesis spaces. The interpretation of Ockham's Razor proposed by Murphy and Pazzani appears rather similar to that of the Akaike Information Criterion (Akaike 1973, Forster and Sober 1994) and may well lead to comparable results. At this stage, very few empirical contradictions of Ockham's Razor have been proposed - however, they have been sufficient to put the validity of Ockham's Razor into question. It is our belief that before any conclusions can be reached over the validity of Ockham's Razor significantly more empirical experimentation will be required. A theoretical proof of our interpretation of Ockham's Razor is currently unavailable, but many well lie in the realms of complexity, information and probability theory. While this investigation would appear to be the first explicit investigation of the message length interpretation of Ockham's Razor, the strength of the MML inference techniques promise much potential for future work.

Acknowledgements

We thank Murray Jorgensen for early discussions on the work of Murphy and Pazzani and on the work of Webb.

References

H. Akaike, *Information theory and an extension of the maximum likelihood principle*, Proceedings of the 2nd International Symposium on Information Theory, (1973), 267-281.

J.M. Bernardo and A.F.M. Smith, *Bayesian theory*, Wiley, Chichester, 1994.

P. Domingos, *The role of Occam's razor in knowledge discovery*, Proceedings of KDD, 1999, 1-19.

D.L. Dowe, G.E. Farr, A.J. Hurst and K.L. Lentin, *Information-theoretic football tipping*, TR. #96/297, Dept Computer Science, Monash University, Melbourne, 11pp, 1996.

M.R. Forster and E. Sober, *How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions*, British Journal for the Philosophy of Science **45** (1994), 1-35.

I.J. Good, *Corroboration, explanation, evolving probability, simplicity, and a sharpened razor*, British Journal Philosophy of Science **19** (1968), 123-143.

W.H. Jefferys and J.O. Berger, *Sharpening Ockham's razor on a Bayesian strop*, TR. 91-44C, Dept. of Statistics, Purdue University, August 1991, 13 pages.

D.V. Lindley, *Bayesian statistics, a review*, SIAM (Philadelphia, PA), 1972, p. 71.

P.M. Murphy, *An empirical analysis of the benefit of decision tree size biases as a function of concept distribution.*, TR. 95-29, Department of Information and Computer Science, University of California, Irvine, 1995.

P.M. Murphy and M.J. Pazzani, *Exploring the decision forest: An empirical investigation of Occam's razor in decision tree induction*, Journal of Artificial Intelligence Research **1** (1994), 257-275.

J.R. Quinlan and R.L. Rivest, *Inferring decision trees using the minimum description length principle*, Information and Computation **80** (1989), 227-248.

J.J. Rissanen, *Modeling by shortest data description*, Automatica **14** (1978), 465-471.

R.J. Solomonoff, *A formal theory of inductive inference*, Information and Control **7** (1964), 1-22,224-254.

R.J. Solomonoff, *Two kinds of probabilistic induction, Computer Journal (Special issue on Kolmogorov Complexity)* **42(4)** (1999), 256-259.

W.T.B. Uther and M.M. Veloso, *The lumberjack algorithm for learning linked decision forests*, Proceedings of PRICAI, 2000, 156-166.

C.S. Wallace and D.M. Boulton, *An information measure for classification*, Computer Journal **11** (1968), 185-194.

C.S. Wallace and D.L. Dowe, *Minimum message length and Kolmogorov complexity*, Computer Journal (Special issue on Kolmogorov Complexity) **42(4)** (1999), 270-283.

C.S. Wallace and D.L. Dowe, *MML clustering of multistate, Poisson, von Mises circular and Gaussian distributions*, Journal of Statistics and Computing **10** (2000), 73-83.

C.S. Wallace and P.R. Freeman, *Estimation and inference by compact coding*, Journal Royal Statistical Society (Series B) **49** (1987), 240-252.

C.S. Wallace and J.D. Patrick, *Coding decision trees*, Machine Learning **11** (1993), 7-22.

G.I. Webb, *Further experimental evidence against the utility of Occam's razor*, Journal of Artificial Intelligence Research **4** (1996), 397-417.