MML Inference of Oblique Decision Trees

Peter J. Tan and David L. Dowe

School of Computer Science and Software Engineering, Monash University,
Clayton, Vic 3800, Australia
ptan@bruce.csse.monash.edu.au

Abstract. We propose a multivariate decision tree inference scheme by using the minimum message length (MML) principle (Wallace and Boulton, 1968; Wallace and Dowe, 1999). The scheme uses MML coding as an objective (goodness-of-fit) function on model selection and searches with a simple evolution strategy. We test our multivariate tree inference scheme on UCI machine learning repository data sets and compare with the decision tree programs C4.5 and C5. The preliminary results show that on average and on most data-sets, MML oblique trees clearly perform better than both C4.5 and C5 on both "right" / "wrong" accuracy and probabilistic prediction - and with smaller trees, i.e., less leaf nodes.

1 Introduction

While there are a number of excellent decision tree learning algorithms such as CART [2], C4.5 and C5 [13], much research effort has been continuously directed to finding new and improved tree induction algorithms. Most decision tree algorithms only test on one attribute at internal nodes, and these are often referred to as univariate trees. One of the obvious limitations of univariate trees is that their internal nodes can only separate the data with hyperplanes perpendicular to the co-ordinate axes. Multivariate decision tree algorithms attempt to generate decision trees by employing discriminant functions at internal nodes with more than one attribute, enabling them to partition the instance space with hyperplanes of arbitrary slope - rather than only parallel to the co-ordinate axes.

We propose an oblique decision tree inference scheme by using the minimum message length (MML) principle [19, 21, 20, 17]. Test results show our new oblique decision tree inference algorithms find smaller trees with better (or near identical) accuracy compared to the standard univariate schemes, C4.5 and C5.

2 MML Inference of Multivariate Decision Trees

MML inference [19, 21, 8, 20, 17, 4, 5, 18] has been successfully implemented in [22] to infer univariate decision trees (refining [14]) and in [12, 16, 17] to infer univariate decision graphs, with the most recent decision graphs [16, 17] clearly out-performing *both* C4.5 and C5 [13] on *both* real-world and artificial data-sets on a range of test criteria - we had better "right" / "wrong" accuracy, substantially

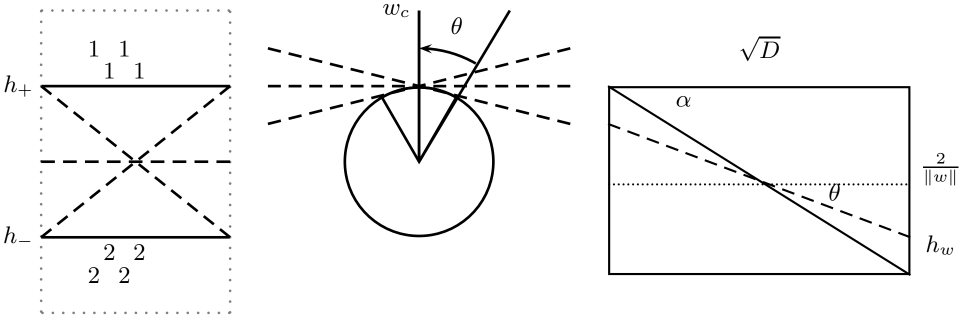

Fig. 1. The set of hyperplanes (Fig. 1a) defined by vector $w \in \Lambda(\theta)$, (Fig. 1b) a partial sphere of radius θ formed by $w \in \Lambda(\theta)$ and (Fig. 1c) the upper bound of θ

better probabilistic score and [17, Table 4] fewer leaf nodes. In this paper, we use MML to infer multivariate decision trees. The new multivariate, oblique, decision tree scheme proposed here generalizes earlier MML decision tree work and re-uses the Wallace and Patrick decision tree coding [22] as part of its coding scheme. For further implementation details, please see [16].

2.1 Encoding an internal split using a linear discriminant function

To infer oblique decision trees by the MML principle, we extend the Wallace and Patrick decision tree coding scheme [22]. The new MML decision tree coding scheme is able to encode an internal split using a linear discriminant function. Firstly, the data falling at an internal node is scaled and normalized so that every data item falls within a D-dimensional unit hyper-cube, where D is the number of input attributes. A linear decision function d(w, x, b)=0 is written as $(\sum_{i=1}^{D} w_i x_i) + b = w \cdot x + b = 0$ where $w, x \in R^D$, $\cdot$ denotes the dot (or scalar) product, and the scalar b is often called the bias. The data is divided into two mutually exclusive sets by the following rules:

If $d(w, x_j, b) > 0, j \in [1, N]$, then x_j is assigned to set I (denoted '1' or '+').

If $d(w, x_j, b) < 0, j \in [1, N]$, then x_j is assigned to set II (denoted '2' or '$-$').

To encode the hyperplane is equivalent to transmitting the vector w and the bias b. Suppose the desired value of the vector w is w_c. If we state w_c exactly (to infinite precision), it will cost infinitely many bits of information in the first part of the message. So instead, we attempt to state a set of vectors $\Lambda(\theta), \theta \in (0, \frac{\pi}{2})$, which is defined as $\Lambda(\theta) = \{w : \arccos(\frac{w \cdot w_c}{\|w\| \cdot \|w_c\|}) < \theta\}$. This is the set of vectors which form an angle less than θ with the optimal vector w_c as illustrated in Fig. 1b. The probability that a randomly picked vector falls into the set is given by $\frac{V_\theta}{V_T}$, where V_θ is the volume of a partial sphere of radius θ and V_T is the total volume of the unit sphere. The value of $\frac{V_\theta}{V_T}$ is given [15] by $(\sin\theta)^{2(D-1)}$, so the information required to specify the set of the vectors is $-\log((\sin\theta)^{2(D-1)})$.

By specifying one data point on each side of the hyperplane h_c, two hyperplanes which are parallel to the decision surface d(w,x,b)=0 are also defined. We denote these two hyperplanes as h_+ and h_-. These (h_+ and h_-) and the other boundaries of the unit cube form a hyper-rectangle as shown in Fig. 1a.

We want to work out the value of θ so that the hyperplanes specified by vectors in the set $\Lambda(\theta)$ do not intersect with the hyperplanes h_+ and h_-. We can imagine a rectangle whose length of one side is the distance between h_+ and h_- and whose length of the other side is $\sqrt{D}$, which is the longest diagonal in a D-dimensional unit cube. As {x: kwx+kb=0} $\equiv$ {x: wx+b=0} for any non-zero k, we can choose w so that the margin between h_+ and h_- is equivalent to $\frac{2}{\|w\|}$. As shown in Figure 1c, given the margin $\frac{2}{\|w\|}$, if $\theta < \alpha$, where $\alpha = \arcsin(\frac{2}{\sqrt{D\|w\|^2+4}})$, one can show that the hyperplane h_w defined by the vector w does not intersect with hyperplanes h_+ and h_- within the D-dimensional hyper-cube (from Fig. 1a).

2.2 Search for the optimal hyperplane

In order to perform faster searches for optimal multivariate splits, we do not use the search heuristic used in OC1 [10] and SADT [9]. Instead, we implement a simple evolution strategy as the preliminary search heuristic for our scheme. A similar approach has appeared in [3], in which promising results were reported. The search process in our scheme can be summarized as follows. Assuming the linear discriminant function in our scheme takes the form $\sum_{i=1}^{d} w_i x_i < w_{d+1}$, for each leaf node L, let M(unsplit) denote the message length of the node L while the node is unsplit, and let M(T) denote the message length of the subtree when node L is split by vector w^T at round T. The algorithm searches for the best vector w via the following steps: Set T=0, input R, MaxP, M(unsplit)

```
1. Re-scale the coefficients of the vector w such that $\sum_{i=1}^{d} w_i^2 = 1$.
2. With $v \sim N(0,1)$, randomly pick $j \in [1, d+1]$, $w_j^{T+1} = w_j^T + v$.
3. if $M(T+1) < M(T)$, go to step 5
4. $w_j^{T+1} = w_j^T$
5. T=T+1; if $T < R$, go to step 1.
6. Randomly pick d (in this paper, d is limited to 2 or 3) attributes
7. P=P+1; if $P < MaxP$, go to step 1
8. if $M(R) < M(unsplit)$, return w, M(R), else return null and M(unsplit).
```

The search process (from steps 2 and 6) is non-deterministic, thus our algorithm is able to generate many different trees. As such, our algorithm can be extended to take advantage of this by choosing the best one (i.e., MML tree) among these trees or by averaging [20, p281] results from these trees.

3 Experiments

3.1 Comparing and scoring probabilistic predictions

To evaluate our new oblique decision tree scheme, we run experiments on nine data sets selected from the UCI Repository [1]. The performance of our scheme is compared with those of C4.5 and C5 [13]. In addition to the traditional right/wrong accuracy, we are also keen to compare the probabilistic performance

[17, sec 5.1] $[7, 6, 11, 16]$ of the learning algorithms. In a lot of domains, like oncological and other medical data, not only the class predictions but also the probability associated with each class is essential. In some domains, like finance, (long term) strategies heavily rely on accurate probabilistic predictions. For C4.5, C5 and our approach, we ascribe class probabilities from frequency counts in leaves using "+1.0" (Laplace estimation) from [17, sec. 5.1]. To compare probabilistic prediction performances, we propose a metric called the related (test data) code length (RCL), defined as $RCL = -\frac{\sum_{i=1}^{n}\log(p_i)}{n\log(M)}$, where n is the total number of test data, M is the arity of the target attribute and p_i is the probability assigned to the real class associated with the test instance i by the model. The related test data code length (RCL) is equivalent to the code length of the test data encoded by a model divided by the code length encoded by the null theory; thus normalizing [17, Sec. 5.1] $[7, 6, 11, 16]$ $-\sum_{i=1}^{n}\log(p_i)$. The smaller RCL, the better the model's performance on probabilistic prediction.

3.2 Data sets

The purpose of the experiment is to have our algorithms perform on real world data, especially on oncological and medical data, such as **Bupa**, **Breast Cancer**, **Wisconsin**, **Lung Cancer**, and **Cleveland**. The nine UCI Repository [1] data-sets used are these five, **Balance**, **Credit**, **Sonar** and **Wine**. For each of the nine data sets, 100 independent tests were done by randomly sampling 90% of the data as training data and testing on the remaining 10%.

4 Discussion

We compare the MML oblique tree scheme to C4.5 and C5. The results from Table 1 clearly suggest that the MML oblique trees are much smaller (fewer leaves) than the C4.5 and C5 univariate trees. The MML oblique trees perform significantly better than C4.5 and C5 (which often have RCL scores worse than the default "random null" of 1.0) on all data-sets. MML oblique trees also have higher "right"/"wrong" accuracy than C4.5 and C5 except (for very close results) on the Bupa and Wine (and Cleveland) data, suggesting a possible need to refine the searches. As expected, none of the algorithms have good results on the Lung Cancer data - learning from a small set of data with a great number of attributes remains a great challenge for machine learning algorithms.

5 Conclusion and Future Research

We have introduced a new oblique decision tree inference scheme by using the MML principle. Our preliminary algorithm produces very small trees with excellent performance on both "right"/"wrong" accuracy and probabilistic prediction. The search heuristic could be (further) improved. Also, as pointed out in section 2.2, the performance of the system may be enhanced by using multiple tree averaging. Further down the track, to use MML coding for internal nodes with SVMs or nonlinear splits is also an interesting research topic, as is generalising oblique trees to oblique graphs. We also wish to apply Dowe's notion of inverse learning

Table 1. Test Results

Name	Metric	C4.5	C5	MML Oblique Tree	Random NULL
Balance	Accuracy(%)	77.8 ±4.3	77.8 ± 4.5	88.5 ± 4.0	33.3
	RCL	0.93±0.12	0.92 ±0.11	0.33 ± 0.08	1.00
	Tree Size	81.6±9.7	41.7 ±4.6	10.4 ±0.9	1
Bupa	Accuracy(%)	65.5 ± 7.4	65.5 ± 7.8	65.1 ± 8.1	50.0
	RCL	1.07±0.22	1.07 ±0.21	0.96 ± 0.15	1.00
	Tree Size	49.2±9.8	27.3 ±5.4	6.7 ± 2.6	1
Breast Cancer	Accuracy(%)	71.2 ± 8.7	71.1 ± 8.4	72.8 ± 8.0	50.0
	RCL	0.88±0.17	0.88 ±0.17	0.84 ± 0.14	1.00
	Tree Size	24.2±8.3	13.1 ±4.2	3.0 ± 0.6	1
Wisconsin	Accuracy(%)	94.6 ± 2.5	94.8 ± 2.5	96.0 ± 2.3	50.0
	RCL	0.26±0.10	0.25 ±0.12	0.21 ± 0.10	1.00
	Tree Size	23.7±5.3	12.3 ±2.8	5.5 ± 0.9	1
Credit	Accuracy(%)	73.2 ± 4.3	73.3 ± 3.8	75.4 ± 4.7	50.0
	RCL	0.88±0.08	0.88 ±0.08	0.79 ± 0.09	1.00
	Tree Size	151.4±17.7	77.6 ±9.1	6.5 ± 2.4	1
Lung Cancer	Accuracy(%)	40.0 ± 23.3	40.7 ± 24.8	46.8 ± 22.4	33.3
	RCL	1.83±0.50	1.86 ±0.65	0.94 ± 0.30	1.00
	Tree Size	12.2±2.3	6.6± 1.1	2.2 ± 0.4	1
Cleveland	Accuracy(%)	77.1 ± 7.6	77.2 ± 7.9	77.2 ± 7.8	50.0
	RCL	0.80±0.24	0.81 ±0.21	0.76 ± 0.22	1.00
	Tree Size	36.7±7.2	20.0 ±4.2	7.3 ± 1.8	1
Sonar	Accuracy(%)	72.8 ± 9.2	73.9 ± 10.0	76.0 ± 9.2	50.0
	RCL	1.07±0.37	1.06 ±0.42	0.98 ± 0.33	1.00
	Tree Size	28.2±3.1	14.9 ±1.6	11.6 ± 9.3	1
Wine	Accuracy(%)	93.6 ± 5.7	93.2 ± 5.8	93.2 ± 6.1	33.3
	RCL	0.42±0.30	0.44 ±0.29	0.28 ± 0.18	1.00
	Tree Size	9.6±1.3	5.4 ±0.7	3.6 ± 0.5	1

[8] and its special case of generalised Bayesian networks [4, 5] to Dowe's notion of a(n inverse) decision graph model where two values of the target attribute have the same probability ratio in every leaf - e.g., the ternary target attribute has values (i) Female, (ii) Male whose height rounds to an even number of cm and (iii) Males whose height rounds to an odd number of cm.

The second author expresses great fondness and gratitude to his mother. We also thank our mentor, Chris Wallace (1933-2004), a quietly-achieving humble unsung genius and brilliant light clearly ahead of his time. Read his works (e.g., via www.csse.monash.edu.au/∼dld/CSWallacePublications/).

References

1. C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998. http://www.ics.uci.edu/∼mlearn/MLRepository.html.
2. Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification And Regression Trees*. Wadsworth & Brooks, 1984.

3. Erick Cantu-Paz and Chandrika Kamath. Using evolutionary algorithms to induce oblique decision trees. In *Proc.Genetic and Evolutionary Computation Conference*, pages 1053–1060, Las Vegas, Nevada, USA, 2000. Morgan Kaufmann.
4. Joshua W. Comley and David L. Dowe. Generalised Bayesian networks and asymmetric languages. In *Proc. Hawaii International Conference on Statistics and Related Fields*, 5-8 June 2003.
5. Joshua W. Comley and David L. Dowe. Minimum message length, MDL and generalised Bayesian networks with asymmetric languages. In P. Grünwald, M. A. Pitt, and I. J. Myung, editors, *Advances in Minimum Description Length: Theory and Applications (MDL Handbook)*. M.I.T. Press, to appear.
6. D.L. Dowe, G.E. Farr, A.J. Hurst, and K.L. Lentin. Information-theoretic football tipping. In N. de Mestre, editor, *Third Australian Conference on Mathematics and Computers in Sport*, pages 233–241. Bond University, Qld, Australia, 1996. http://www.csse.monash.edu.au/∼footy.
7. D.L. Dowe and N. Krusel. A decision tree model of bushfire activity. In *(Technical report 93/190) Dept. Comp. Sci., Monash Uni., Clayton, Australia*, 1993.
8. D.L. Dowe and C.S. Wallace. Kolmogorov complexity, minimum message length and inverse learning. In *14th Australian Statistical Conference (ASC-14)*, page 144, Gold Coast, Qld, Australia, 6-10 July 1998.
9. David G. Heath, Simon Kasif, and Steven Salzberg. Induction of oblique decision trees. In *International Joint Conference on AI (IJCAI)*, pages 1002–1007, 1993.
10. Sreerama K. Murthy. *On Growing Better Decision Trees from Data*. PhD thesis, The John Hopkins University, 1997.
11. S.L. Needham and D.L. Dowe. Message length as an effective Ockham's razor in decision tree induction. In *Proc. 8th International Workshop on Artificial Intelligence and Statistics*, pages 253–260, Key West, Florida, U.S.A., Jan. 2001.
12. J.J. Oliver and C.S. Wallace. Inferring Decision Graphs. In *Workshop 8 International Joint Conference on AI (IJCAI)*, Sydney, Australia, August 1991.
13. J.R. Quinlan. *C4.5 : Programs for Machine Learning*. Morgan Kaufmann,San Mateo,CA, 1992. The latest version of C5 is available from http://www.rulequest.com.
14. J.R. Quinlan and R. Rivest. Inferring Decision Trees Using the Minimum Description Length Principle. *Information and Computation*, 80:227–248, 1989.
15. R. Schack, G. M. D. Ariano, and C. M. Caves. Hypersensitivity to perturbation in the quantum kicked top. *Physical Review E.*, 50:972–987, 1994.
16. P.J. Tan and D.L. Dowe. MML inference of decision graphs with multi-way joins. In *Proc. 15th Australian Joint Conf. on AI, LNAI 2557 (Springer)*, pages 131–142, Canberra, Australia, 2-6 Dec. 2002.
17. P.J. Tan and D.L. Dowe. MML inference of decision graphs with multi-way joins and dynamic attributes. In *Proc. 16th Australian Joint Conf. on AI, LNAI 2903 (Springer)*, pages 269–281, Perth, Australia, Dec. 2003. http://www.csse.monash.edu.au/∼dld/Publications/2003/Tan+Dowe2003.ref .
18. Chris Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Springer, to appear.
19. C.S. Wallace and D.M. Boulton. An Information Measure for Classification. *Computer Journal*, 11:185–194, 1968.
20. C.S. Wallace and D.L. Dowe. Minimum Message Length and Kolmogorov Complexity. *Computer Journal*, 42(4):270–283, 1999.
21. C.S. Wallace and P.R. Freeman. Estimation and Inference by Compact Coding. *Journal of the Royal Statistical Society. Series B*, 49(3):240–265, 1987.
22. C.S Wallace and J.D. Patrick. Coding Decision Trees. *Machine Learning*, 11:7–22, 1993.