Bayesian Model Averaging in Rule Induction

Pedro Domingos*
Department of Information and Computer Science
University of California, Irvine
Irvine, California 92697, U.S.A.
pedrod@ics.uci.edu
http://www.ics.uci.edu/~pedrod

Abstract

Bayesian model averaging (BMA) can be seen as the optimal approach to any induction task. It can reduce error by accounting for model uncertainty in a principled way, and its usefulness in several areas has been empirically verified. However, few attempts to apply it to rule induction have been made. This paper reports a series of experiments designed to test the utility of BMA in this field. BMA is applied to combining multiple rule sets learned from different subsets of the training data, to combining multiple rules covering a test example, to inducing technical rules for foreign exchange trading, and to inducing conjunctive concepts. In the first two cases, BMA is observed to produce lower accuracies than the *ad hoc* methods it is compared with. In the last two cases, BMA is observed to typically produce the same result as simply using the best (maximum-likelihood) rule, even though averaging is performed over all possible rules in the space, the domains are highly noisy, and the samples are medium- to small-sized. In all cases, this is observed to be due to BMA's consistent tendency to assign highly asymmetric weights to different models, even when their accuracy differs by little, with most models (often all but one) effectively having no influence on the outcome. Thus the effective number of models being averaged is much smaller for BMA than for common *ad hoc* methods, leading to a smaller reduction in variance. This suggests that the success of the multiple models approach to rule induction is primarily due to this variance reduction, and not to its being a closer approximation to the Bayesian ideal.

Keywords: Predictive modeling, classification, rule induction, model uncertainty, multiple models, Bayesian methods.

*Partly supported by JNICT/PRAXIS XXI and NATO scholarships. Thanks also to Dennis Kibler.

1 Introduction

Most approaches to predictive modeling assume that there is only one "right" model in the model space under consideration, and accordingly proceed to seek and use only the "best" one. However, in practice it is seldom the case that a single "right" model can be unequivocally identified from the data. When multiple plausible models exist, ignoring all but one will in general be suboptimal with respect to the goal of minimizing error. Combining multiple models in some way then becomes an attractive alternative, and it has recently received much attention (e.g., [5]). Most combination methods in the literature are *ad hoc* in the sense that they have been empirically observed to work well, but have no firm theoretical foundation. The Bayesian approach [2] provides such a foundation, and the hope is that closer adherence to it will produce improved results over more heuristic methods. In essence, Bayesian model averaging (BMA) weights each model by its posterior probability (i.e., its probability given the observed data). In areas such as regression and discrete graphical models, it has been verified to produce improvements over using the single most likely model [13]).

Applications of BMA to machine learning methods have been sparser. One of the main difficulties is that, in most cases, the number of possible models is far too large to be exhaustively considered, and no closed form for the relevant sums (or integrals) is known. A plausible approximation is then to only attempt to find several of the most probable models, and average over these. Buntine [4] has successfully applied this approach to decision tree induction (see also [16]). Applications of BMA to rule induction have been carried out by Kononenko [10] and by Ali and Pazzani [1]. The latter found that it often improved accuracy relative to using the single "best" rule set. In this paper we compare BMA with *ad hoc* methods for combining multiple rule sets.

2 Bayesian Averaging of Rule Sets

RISE [7] is a rule induction system that assigns each test example to the class of the nearest rule according to a similarity measure, and thus implicitly partitions the instance space into the regions won by each of the rules. Its learning time on large databases can be much reduced by randomly dividing the database into several smaller ones, running the system on each one separately, and combining the results. This combination was originally performed by letting each rule set vote for the class it predicts, with a weight equal to the training-set accuracy[1] of the rule that won the example in that rule set [8]. This *ad hoc* approach was compared with BMA on eight of the larger databases in the UCI repository[14][2]. BMA was applied in the following form, very similar to that of [4] and [1].

Let n be the sample size, $\vec{x}$ the training examples, $\vec{c}$ the corresponding class labels, and H the set of models induced (i.e., each element h of H is a rule set). Then, by Bayes's Theorem, and assuming the examples are drawn independently:

[1] With the Laplace correction [15].

[2] Credit, diabetes, annealing, chess, hypothyroid, splice junctions, mushroom and shuttle.

$$Pr(h|\vec{x},\vec{c}) = \frac{Pr(h)}{Pr(\vec{x},\vec{c})} \prod_{i=1}^{n} Pr(x_i, c_i|h) \quad (1)$$

where the data prior $Pr(\vec{x},\vec{c})$ is the same for all models, and can be ignored. Note that each model is induced from a different subdatabase, so, strictly speaking, $\vec{x}$, $\vec{c}$ and their components should be indexed by h. We have omitted this for the sake of simplicity. $Pr(h)$ is the prior probability of h, and is assumed uniform (i.e., Dirichlet with parameter $\alpha = 1$). For each pair (x_i, c_i) in the training set, $Pr(x_i, c_i|h)$ is computed as the probability of an example having class c_i given that it is in the region won by the rule that wins x_i. This probability is estimated empirically from the examples won by that rule. Let r be this rule, n_r the total number of examples it wins, and n_{r,c_i} the number of examples of class c_i that it wins. Then:

$$\hat{Pr}(x_i, c_i|h) = \frac{n_{r,c_i}}{n_r} \quad (2)$$

This is analogous to the treatment in [4], using the partition induced by the rules in the same way [4] uses the partition induced by a decision tree. Finally, a test example x is assigned to the class that maximizes:

$$Pr(c|x, H) = \sum_{h \in H} Pr(c|x, h)\, Pr(h|\vec{x},\vec{c}) \quad (3)$$

With subdatabases of 100 examples each, BMA produced lower accuracy than the *ad hoc* method in every domain.[3] With subdatabases of size 500 (and therefore fewer subdatabases), BMA was more accurate in one domain, as accurate in two, and less accurate in five. Inspection of the posterior probabilities of the rule sets revealed that, in most cases, a single rule set had a higher posterior than the sum of all the others, leading the ensemble to often behave as that rule set by itself.[4]

3 Bayesian Averaging of Individual Rules

When rules of different classes cover a test example, some procedure for deciding the outcome is necessary. Recent versions of the CN2 algorithm [6] let each rule vote for each class with the number of examples of that class it covers. C4.5RULES [17] gives precedence to rules of the class for which the fewest false positives are produced. ITRULE [18] assumes all the rule left-hand sides are independent of each other given the class, and thus makes direct use of Bayes' theorem.

BMA can potentially be applied to this problem, since rules can be regarded as alternative models for the region where they intersect. This is an unusual approach, in that it treats each such region in the way that BMA traditionally treats the whole observation space, but the fact that each region will typically have few examples compared to the whole space means that applying BMA here may be correspondingly more useful. A significant difficulty is that

[3] Applying the Laplace correction to the probability estimates used made no difference.

[4] The more asymmetric the class distribution given the winning rule in that set, the more likely this is to occur. RISE and most rule induction algorithms are designed to maximize this asymmetry.

different rules generally cover different numbers of examples, and these will be differently distributed, making it impossible to ignore the data prior probabilities (the denominator in Bayes' theorem). Considering only training examples covered by all the conflicting rules is not viable, because there are often very few, or none.

Three different approaches for Bayesian averaging of individual rules were implemented. All assumed a uniform prior among rules. The first approach was based on directly computing the data priors, estimating each class prior $Pr(c_i)$ from the whole sample:

$$Pr(r|\vec{x},\vec{c}) \propto \prod_{x_i \text{ won by } r} \frac{Pr(c_i|r)}{Pr(c_i)} \tag{4}$$

The second approach was based on noting that each rule partitions the observation space into the region that it wins and its complement, and computing the probability of the observed class distribution among the two. Thus the data is the whole sample, and therefore the data prior is the same for all rules and can be ignored. Let $y_i = 1$ if r wins x_i and 0 otherwise, and $\overline{r}$ represent the negation of r (strictly speaking, of its antecedent). Then:

$$Pr(r|\vec{y},\vec{c}) \propto \prod_{y_i=1} Pr(c_i|r) \cdot \prod_{y_i=0} Pr(c_i|\overline{r}) \tag{5}$$

The third approach was based on considering the data to be, not the examples and their classes *per se*, but the triplet $(n_r, \vec{e_r}, \vec{e_{\overline{r}}})$, where n_r is the total number of examples rule r wins, and the jth component of $\vec{e_r}$ ($\vec{e_{\overline{r}}}$) is the number of examples of the jth class it wins (does not win). n_r is treated as having a binomial distribution, $\vec{e_r}$ and $\vec{e_{\overline{r}}}$ as having multinomial distributions (conditioned on n_r and, for $\vec{e_{\overline{r}}}$, on $\vec{e_r}$). Once again, the data prior is the same for all rules, and can be ignored.

Each approach was applied to the rule sets produced by CN2 and C4.5RULES on 25 datasets from the UCI repository, and compared with the system's native rule combination scheme. In each case, the BMA approach was less accurate than the *ad hoc* scheme in a large majority of the datasets (wins-ties-losses: 1-3-21, 4-1-20, 8-3-14 with C4.5RULES; 5-2-18, 2-3-20, 4-1-20 with CN2). Compared with CN2's weighting scheme, BMA produced far more asymmetric rule weights. Compared with C4.5RULES, the BMA scheme that produced the least asymmetric weights (the third) fared best, and the one producing the most asymmetric ones (the first) fared worst.

4 Bayesian Averaging of Foreign Exchange Trading Rules

In each of the previous cases, BMA could not be applied in its ideal form, due to the very large number of possible models. However, this will be feasible in sufficiently restricted model spaces. One significant application where these arise is foreign exchange prediction, where the goal is to maximize the return from investing in a foreign currency, by predicting whether it will rise or fall against the US dollar. An approach that is used by some traders, and that has been validated by large-scale empirical studies [11], involves the use of so-called *technical rules* of the form "If the s-day moving average of the currency's exchange rate rises above

the t-day one, buy; else sell." The choice of s and t, with $t > s$, can be made empirically. If a maximum value t_{max} is set for t (and, in practice, moving averages of more than a few months are never considered), the total number of possible rules is $t_{max}(t_{max}-1)/2$. It is thus possible to compare the return yielded by the single most accurate rule with that yielded by averaging *all* possible rules according to their posterior probabilities. These are computed assuming a uniform prior on rules/hypotheses and ignoring the data prior (see Equation 1):

$$Pr(h|\vec{x},\vec{c}) \propto \prod_{i=1}^{n} Pr(x_i, c_i|h) \tag{6}$$

Let the two classes be + (rise/buy) and − (fall/sell). $Pr(x_i, c_i|h)$ can take only four values: $Pr(+|+)$, $Pr(-|+)$, $Pr(+|-)$ and $Pr(-|-)$. Let n_{-+} be the number of examples in the sample which are of class − but for which rule h predicts +, and similarly for the other combinations. Let $n_{\cdot+}$ be the total number of examples for which h predicts +, and similarly for $n_{\cdot-}$. Then, estimating probabilities from the sample as in Equation 2:

$$\hat{Pr}(h|\vec{x},\vec{c}) \propto \left(\frac{n_{++}}{n_{\cdot+}}\right)^{n_{++}} \left(\frac{n_{-+}}{n_{\cdot+}}\right)^{n_{-+}} \left(\frac{n_{+-}}{n_{\cdot-}}\right)^{n_{+-}} \left(\frac{n_{--}}{n_{\cdot-}}\right)^{n_{--}} \tag{7}$$

A comparison of this approach with the single most accurate rule was carried out using daily data on five currencies for the years 1973-87, from the Chicago Mercantile Exchange [19]. The first ten years were used for training (2341 examples) and the remainder for testing. The fact that the domain is extremely noisy (typical accuracies are only slightly above 50%), and that no rule can claim to be the "right" model, favors the use of BMA. However, in three cases (German mark, British pound and Swiss franc) BMA produced exactly the same results as the single best rule, and in a fourth case (Canadian dollar) the results were very similar (slightly worse).[5] Inspection of the posteriors showed this to be due in each case to the presence of a dominant peak in the (s,t) plane, in spite of the high level of noise.

5 Bayesian Averaging of Conjunctions

Because the results of the previous section might be specific to the foreign exchange domain, the following experiment was carried out using artificially generated Boolean domains. Classes were assigned at random to examples described by a features. All conjunctions of 3 of those features were then generated (a total of $a(a-1)(a-2)/6$), and their posterior probabilities were estimated from a random sample composed of half the possible examples. The experiment was repeated for $a = 7, 8, 9, \ldots, 15$. Because the class was random, the accuracy of both BMA and the best conjunction[6] was always approximately 50%. However, even in this situation of pure noise and no possible "right" conjunction of 3 features, the posterior distributions were still highly asymmetric (e.g., the average posterior excluding the

[5] In the fifth case (Japanese yen) BMA chose to hold USD throughout, leading to zero return. This was due to downward movements being in the majority for all rules both when the rule held and did not, even though on average the yen went up, and points out the limitations of only making binary predictions. The single best rule, however, produced a return of 84% in five years.

[6] Predicting the class with highest probability given that the conjunction is satisfied when it is (estimated from the sample), and similarly when it is not.

maximum was 14% of the maximum for $a = 7$, and decreased to 6% for $a = 13$). As a result, BMA still made the same prediction as the "best" conjunction 83.9% of the time for $a = 7$, decreasing to 64.4% for $a = 13$.[7]

6 Discussion

It is well known that, in some applications, the model with highest posterior probability dominates all others, making BMA equivalent to approaches that simply pick that model [12]. However, the results described in this paper are still surprising, and require interpretation. If multiple models derive their power from being a closer approximation of the Bayesian ideal, as Buntine [4] suggests, then BMA should produce higher accuracy than *ad hoc* averaging. If BMA is most advantageous compared to the single best model when samples are small and noisy, and when the "right" model is clearly not in the model space considered, then BMA would be expected to outperform the single best rule in domains like the foreign exchange one in Section 4. It should also not give high preference to some models over others when all are in fact equally poor, as in Section 5.

In the Bayesian view, if multiple models are considered but some attain a weight so high that the others are rendered irrelevant, this is as it should be: the models with low posterior are simply not good models, and are properly ignored. An alternative view of multiple models attributes the error reduction they produce to variance reduction [3, 9], and views giving a significant weight to a model as potentially useful, even if that model is not as good as the best. In this view, the highly asymmetric weights produced by BMA are disadvantageous compared to the more even ones typically found in *ad hoc* methods, because BMA's weights reduce the number of models effectively being considered, and thus increase the variance. The results reported in this paper support this view.

In principle, if all possible models are considered, BMA will produce the optimal result for the model space and priors given. However, this ignores that the probabilities BMA uses (apart from the priors) generally have to be estimated from the sample, and thus are themselves sensitive to variance in it. Rule posteriors vary exponentially as the distribution of examples covered becomes more asymmetric, and therefore BMA is highly sensitive to small, random variations in the sample. To see this, consider any two of the conjunctions in Section 5, h_1 and h_2. Using the notation of Section 4, for each conjunction $Pr(+|+) = Pr(-|+) = Pr(+|-) = Pr(-|-) = \frac{1}{2}$, by design. By Equation 6, $Pr(h_1|\vec{x},\vec{c})/Pr(h_2|\vec{x},\vec{c}) = 1$. In other words, given a large enough sample, the two conjunctions should appear approximately equally likely. Now suppose that: $n = 4000$; for conjunction h_1, $n_{++} = n_{--} = 1050$ and $n_{-+} = n_{+-} = 950$; and for conjunction h_2, $n_{++} = n_{--} = 1010$ and $n_{-+} = n_{+-} = 990$. The resulting estimates of $\hat{Pr}(+|+), \ldots, \hat{Pr}(-|-)$ for both conjunctions are quite good; all are within 5 - 1% of the true values. However, the estimated ratio of conjunction posteriors is, by Equation 7:

$$\frac{\hat{Pr}(h_1|\vec{x},\vec{c})}{\hat{Pr}(h_2|\vec{x},\vec{c})} = \frac{\left(\frac{1050}{2000}\right)^{1050}\left(\frac{950}{2000}\right)^{950}\left(\frac{950}{2000}\right)^{950}\left(\frac{1050}{2000}\right)^{1050}}{\left(\frac{1010}{2000}\right)^{1010}\left(\frac{990}{2000}\right)^{990}\left(\frac{990}{2000}\right)^{990}\left(\frac{1010}{2000}\right)^{1010}} \simeq 120$$

[7] This decrease was not due to a flattening of the posteriors as the sample size increased (the opposite occurred), but to the class probabilities given the value of each conjunction converging to the 50% limit.

In other words, even though the two conjunctions should appear similarly likely and have similar weights in the averaging process, h_1 actually has a far greater weight than h_2; enough so, in fact, that Bayes-averaging between h_1 and 100 conjunctions with observed frequencies similar to h_2's is equivalent to always taking only h_1 into account. When many models with similar "true" posteriors are being averaged, the probability of one (or a few) having a significantly skewed distribution of observations purely by chance is quite high; and this model (or these few) will tend to wipe out the influence of all the others.[8] In cases where the "true" posteriors are sufficiently different and the sample is large enough, that difference should prevail over such random effects. However, when the sample is small, the "true" difference between posteriors can very easily be submerged.

Thus the use of BMA with small samples is difficult and error-prone. On the other hand, when the sample is large and the model space is limited, the same exponential variation will tend to make a single model dominate all others, and make BMA equivalent to that single model. In model spaces that allow many different hypotheses to completely fit the training data (e.g., rule sets), all these hypotheses will have equal likelihood. However, because it is not computationally feasible to find all models, preference goes to those that avoid overfitting the data, and not to those with maximum posteriors. Among the former, even small differences in accuracy will lead to large differences in likelihood, again tending to make BMA equivalent to choosing the best model. Because larger model spaces have larger variance, this will again forgo the advantages of multiple models, even with larger samples.

References

[1] K. Ali and M. Pazzani. Classification using Bayes averaging of multiple, relational rule-based models. In D. Fisher and H.-J. Lenz, editors, *Learning from Data: Artificial Intelligence and Statistics V*, pages 207–217. Springer-Verlag, New York, NY, 1996.

[2] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York, NY, 1985.

[3] L. Breiman. Bias, variance and arcing classifiers. Technical Report 460, Statistics Department, University of California at Berkeley, Berkeley, CA, 1996. Available electronically as ftp://ftp.stat.berkeley.edu/users/breiman/arcall.ps.Z.

[4] W. L. Buntine. *A Theory of Learning Classification Rules.* PhD thesis, School of Computing Science, University of Technology, Sydney, Australia, 1990.

[5] P. Chan, S. Stolfo, and D. Wolpert, editors. *Proceedings of the AAAI-96 Workshop on Integrating Multiple Learned Models for Improving and Scaling Machine Learning Algorithms.* AAAI Press, Portland, OR, 1996.

[6] P. Clark and R. Boswell. Rule induction with CN2: Some recent improvements. In *Proceedings of the Sixth European Working Session on Learning*, pages 151–163, Porto, Portugal, 1991. Springer-Verlag.

[8] This might not be true if those other models all voted in the same direction; but in practice this is very unlikely, and they will tend to cancel each other out, further enhancing this effect.

[7] P. Domingos. Unifying instance-based and rule-based induction. *Machine Learning*, 24:141–168, 1996.

[8] P. Domingos. Using partitioning to speed up specific-to-general rule induction. In *Proceedings of the AAAI-96 Workshop on Integrating Multiple Learned Models for Improving and Scaling Machine Learning Algorithms*, Portland, OR, 1996. AAAI Press.

[9] J. H. Friedman. On bias, variance, 0/1 - loss, and the curse-of-dimensionality. Technical report, Department of Statistics and Stanford Linear Accelerator Center, Stanford University, Stanford, CA, 1996. Available electronically as ftp://playfair.stanford.edu/pub/friedman/kdd.ps.Z.

[10] I. Kononenko. Combining decisions of multiple rules. In B. du Boulay and V. Sgurev, editors, *Artificial Intelligence V: Methodology, Systems, Applications*, pages 87–96. Elsevier, Amsterdam, 1992.

[11] B. LeBaron. Technical trading rules and regime shifts in foreign exchange. Technical report, Department of Economics, University of Wisconsin at Madison, Madison, WI, 1991.

[12] D. MacKay. Bayesian interpolation. *Neural Computation*, 4:415–447, 1992.

[13] D. Madigan, A. E. Raftery, C. T. Volinsky, and J. A. Hoeting. Bayesian model averaging. In *Proceedings of the AAAI-96 Workshop on Integrating Multiple Learned Models for Improving and Scaling Machine Learning Algorithms*, Portland, OR, 1996. AAAI Press.

[14] C. J. Merz, P. M. Murphy, and D. W. Aha. UCI repository of machine learning databases. Machine-readable data repository, Department of Information and Computer Science, University of California at Irvine, Irvine, CA, 1996.

[15] T. Niblett. Constructing decision trees in noisy domains. In *Proceedings of the Second European Working Session on Learning*, pages 67–78, Bled, Yugoslavia, 1987. Sigma.

[16] J. J. Oliver and D. J. Hand. On pruning and averaging decision trees. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 430–437, Tahoe City, CA, 1995. Morgan Kaufmann.

[17] J. R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA, 1993.

[18] P. Smyth, R. M. Goodman, and C. Higgins. A hybrid rule-based/Bayesian classifier. In *Proceedings of the Ninth European Conference on Artificial Intelligence*, pages 610–615, Stockholm, Sweden, 1990. Pitman.

[19] A. S. Weigend, B. A. Huberman, and D. E. Rumelhart. Predicting sunspots and exchange rates with connectionist networks. In M. Casdagli and S. Eubank, editors, *Nonlinear Modeling and Forecasting*, pages 395–432. Addison-Wesley, Redwood City, CA, 1992.