Bayesian model averaging is not model combination

Thomas P. Minka

July 6, 2000

In a recent paper, Domingos (2000) compares Bayesian model averaging (BMA) to other model combination methods on some benchmark data sets, is surprised that BMA performs worst, and suggests that BMA may be flawed. These results are actually *not* surprising, especially in light of an earlier paper by Domingos (1997) where it was shown that model combination works by enriching the space of hypotheses, not by approximating a Bayesian model average. And the only flaw with BMA is the belief that it is an algorithm for model combination, when it is not.

Bayesian model averaging is best thought of as a method for 'soft model selection.' It answers the question: "Given that all of the data so far was generated by *exactly one* of the hypotheses, what is the probability of observing the new pair (c, x)?" The soft weights in BMA only reflect a statistical inability to distinguish the hypothesis based on limited data. As more data arrives, the hypotheses become more distinguishable and BMA will *always* focus its weight on the most probable hypothesis, just as the posterior for the mean of a Gaussian focuses ever more narrowly on the sample mean. Mathematically, we can write the BMA rule as

$$p((c,x)|D) \propto \sum_h p((c,x), D|h)p(h) \tag{1}$$

which emphasizes the assumption that *exactly one* hypothesis is responsible for *all* of the data.

A simple example can illustrate the difference between model combination and BMA. Let the true class assignments be as shown in figure 1(a): an instance is in class 'o' iff it is inside at least two of the circles. Let the circles be our three hypotheses. The best way to combine them is with a uniform vote—this gives 100% accuracy. But BMA will not do this; it will focus its weight on the topmost circle, because this circle is the most homogeneous and therefore most likely to have generated the data. (The circle placement is not perfectly symmetric.) Figure 1(b) plots the accuracy on a fixed set of test data, as the training set size increases. As expected, BMA has *worse* performance with more data, because it moves away from the optimal uniform weighting. This happens no matter how similar the hypotheses are in error rate, as long as the error rates are different. Even if the error rates are 20% and 19.99%, BMA will eventually put all weight on the latter hypothesis.

This kind of model mismatch may be to blame in all of Domingos' results. This reminds us that Internet benchmarks in general are not a good way to analyze an algorithm's behavior. Such benchmarks only measure the robustness of an algorithm to the vagaries of real-world data in different domains—they don't measure how well an algorithm exploits the domain assumptions it was designed for.

So now we know that to do model combination, we should not use BMA on the models. What should we do instead? We *can* use Bayesian methods to perform model combination, as long

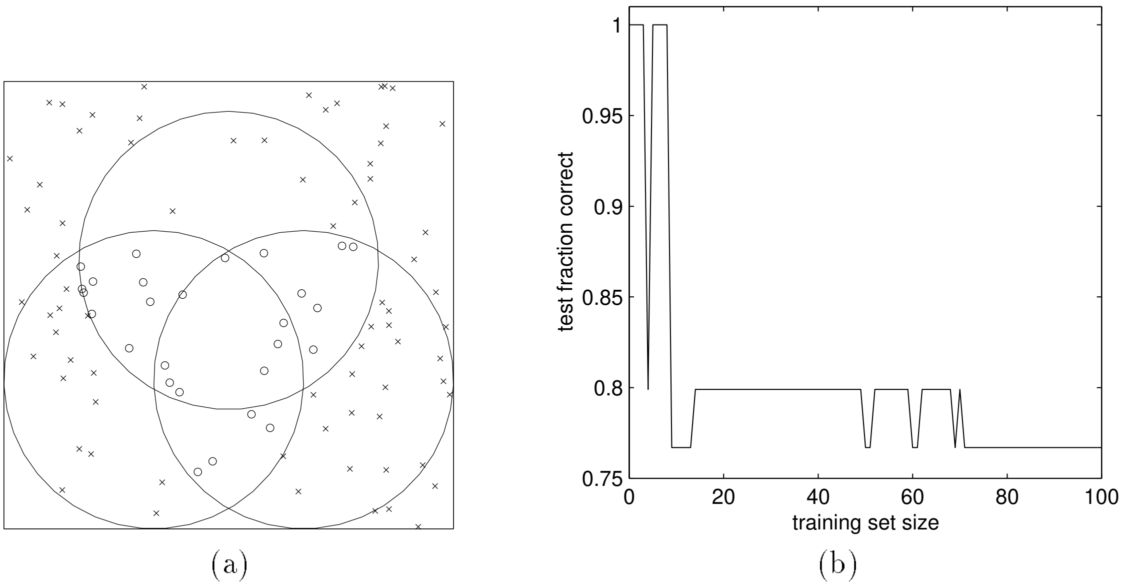

Figure 1: (a) Classification problem. The optimal solution is a uniform vote between the circles. (b) The test-set accuracy of BMA, as a function of training set size. BMA always focuses on the topmost circle, even though the other two circles have nearly the same accuracy.

as we ask the right question. For example: "Given that all of the data so far was generated by *some linear combination* of the hypotheses, what is the probability of observing the new pair (c, x)?" This is BMA applied to a new hypotheses space of "stacked" models. On the circle problem, it will converge to the optimal uniform vote.

References

Domingos, P. (1997). Why does bagging work? A Bayesian account and its implications. *KDD'97*. `http://www.cs.washington.edu/homes/pedrod/kdd97.ps.gz`.

Domingos, P. (2000). Bayesian averaging of classifiers and the overfitting problem. *ICML'00*. `http://www.cs.washington.edu/homes/pedrod/mlc00b.ps.gz`.