Bayesian Averaging of Classifiers and the Overfitting Problem

Rayid Ghani

ML Lunch – 11/13/00

BMA is a form of Ensemble Classification

- Set of Classifiers
- Decisions combined in ”some” way
- Unweighted Voting
  - Bagging, ECOC etc.
- Weighted Voting
  - Weight $\propto$ accuracy (training or holdout set), LSR (weights $\propto$ 1/variance)
  - Boosting

Bayesian Model Averaging

- All possible models in the model space used (weighted by their probability of being the "Correct" model)
- Posterior of a model = Prior * Likelihood given data
- Optimal given the correct model space and priors
- Claimed to obviate the overfitting problem by cancelling the effects of different overfitted models (Buntine 1990)

BMA - Training

posterior prior likelihood

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

ignored noise model

$$P(x_i, c_i \mid h) = P(x_i \mid h) P(c_i \mid x_i, h)$$

$$P(c_i \mid x_i, h) = \begin{cases} 1-\varepsilon & \text{If h predicts correct class } c_i \text{ for } x_i \\ \varepsilon & \text{otherwise} \end{cases}$$

OR

$$P(c_i \mid x_i, h) = \frac{n_{r,c_i}}{n_r}$$

BMA - Testing

$$P(C \mid x, \vec{x}, \vec{c}, H) = \sum_{h \in H} P(c \mid x, h) P(h \mid \vec{x}, \vec{c})$$

Pure Classification Model

P(c|x,h)=1 for the class predicted by h for x

OR

Class Probability Model

Problems

- How to get the priors
- How to get the correct model space
- Model space too large – approximation required
  - Model with highest posterior, Sampling (Imp sampling,MCMC)

BMA of Bagged C4.5 Rules

- Bagging is an approximation of BMA by importance sampling where all samples are weighed equally
- Weighting the models by their posteriors should lead to a better approximation
- Experimental Results
  - Every version of BMA performed worse than bagging on 19 out of 26 UCI datasets
  - Posteriors skewed – dominated by a single rule model – model selection rather than averaging

Experimental Results

- Every version of BMA performed worse than bagging on 19 out of 26 UCI datasets 
- Best performing BMA was uniform class noise and pure classification
- Posteriors skewed – dominated by a single rule model even though error rates were similar
- Model selection rather than averaging?

Bagging as Imp Sampling

- Want to approximate $\sum f(x)p(x) = \sum f(x)\left[\frac{p(x)}{q(x)}\right]q(x)$
- Sample according to q(x) and compute the average of f(x)p(x)/q(x) for points x sampled
- Each sampled value will have weight p(x)/q(x)

BMA of various learners

- RISE Rule sets with partitioning
  - 8 databases from UCI
  - BMA worse than RISE in every domain
- Trading Rules
  - If the *s*-day moving average rises above the *t*-day one, buy; else sell ($t>s$, set t_{max})
  - Intuition (there is no single right rule so BMA should help)
  - BMA AGAIN similar to choosing the single best rule

- Likelihood of a model $(1-\varepsilon)^{s}\varepsilon^{n-s}$ increases exponentially with with s/n
- Small random variation in the sample can sharply increase the likelihood of a model
- Even if a small fraction of terms is considered, the probability of one term being very large purely by chance is very high
- The better the approximation (more terms), the worse the averaging performs?

Overfitting in BMA

- Issue of overfitting is usually ignored (Freund et al. 2000)
- Is overfitting the explanation for the poor performance of BMA?
- Preferring a hypothesis that does not truly have the lowest error of any hypothesis considered, but by chance has the lowest error on training data.
- Overfitting is the result of the likelihood’s exponential sensitivity to random fluctuations in the sample and increases with # of models considered

To BMA or not to BMA?

- Net effect will depend on which effect prevails?
  - Increased overfitting (small if few models are considered)
  - Reduction in error obtained by giving some weight to alternative models (skewed weights => small effect)
- Ali & Pazzani (1996) report good results but bagging wasn't tried
- Domingos (2000) used bootstrapping before BMA so the models were built from less data

Spectrum of ensembles

Overfitting

Bagging

Boosting

BMA

Asymmetry of weights

Bibliography

- Domingos
- Freund, Mansour, Schapire
- Ali, Pazzani