

# Credal Model Averaging: an Extension of Bayesian Model Averaging to Imprecise Probabilities

G. Corani M. Zaffalon

IDSIA  
Switzerland  
`giorgio{zaffalon}@idsia.ch`

ECML PKDD '08

# Outline

- 1 Methodological overview
  - Bayesian model averaging for Naive Bayes.
  - Credal model averaging for Naive Bayes.
- 2 Experiments
- 3 Conclusions

# Model Uncertainty

- Setup: classification using Naive Bayes (NBC).
- Given  $N$  features, we can design  $2^N$  NBCs, each with a different feature set.
- *Model uncertainty*: multiple NBCs, with different feature sets, are consistent with the data; one is not sure which one is the best.
- In this setting, model uncertainty is linked to *feature selection*.
- Considering a single, supposedly best classifier would ignore model uncertainty.
- *Model averaging*: we average over all the  $2^N$  models, instead of choosing a single classifier.



# Bayesian Model Averaging (BMA)

## Probability computation

- Computes a weighted average of the probabilities returned by the different classifier.
- The weight of each classifier is proportional to its posterior probability.

$$P_{BMA}(c_1|\mathbf{d}) = \sum_{M_j} P(c_1|M_j, \mathbf{d})P(M_j|\mathbf{d})$$

- $P_{BMA}(c_1|\mathbf{d})$ : prob. of class  $c_1$  given the data  $\mathbf{d}$ , computed by BMA;
- $P(c_1|M_j, \mathbf{d})$ : prob. of class  $c_1$ , according to classifier  $M_j$ ;
- $P(M_j|\mathbf{d})$ : posterior probability of classifier  $M_j$ .

## BMA for Naive Bayes (Dash & Cooper, 2002)

- Working with  $2^N$  models can be unfeasible (often, approximated solutions are used).
- D&C have designed an algorithm which computes BMA for Naive Bayes *exactly*, without approximations.
- The algorithm of D&C is also efficient, as it eventually implements BMA as a single summary NBC.
- Still, there is no discussion about the sensitivity of BMA to the choice of the prior distribution over the models.

## BMA and prior sensitivity

- The posterior probability of each classifier is given by the product of its *likelihood* and its *prior probability*.
- The prior distribution over the models has to be set by the investigator; this is an open problem for BMA.
- A flat prior is usually adopted, trying to be non-informative.
- However, especially on *small data sets*, different classes can be returned as the most probable one, depending on the prior over the models (*prior-dependent classification*).
- Prior-dependent classifications might be unreliable.
- This problem affects *any* chosen prior.

# Credal model averaging (CMA)

- CMA, currently developed only for Naive Bayes, is designed to address the problem of prior specification in BMA.

## Main ideas

- CMA considers a set of priors over the models (**credal set**) instead of a unique prior.
- The credal set is turned into a set of posteriors over the models via Bayes' rule.
- Eventually, this leads to a set of posteriors over the classes of the problem.
- CMA returns the classes that are *non-dominated* within the set of posterior distributions.

# The credal set

- We consider all the prior distributions which satisfy the condition  $\epsilon < p(M_j) < 1 - \epsilon$  for any model  $M_j$ .
- The credal set lets the prior probability of each model vary between  $\epsilon$  and  $1 - \epsilon$ .
- This hence models *prior ignorance* about the relative credibility of the competing models.
- Remark: a single flat prior can instead be seen as modelling *indifference* between models.

# Test of dominance and indeterminate classifications

## Definition

Class  $c_1$  dominates  $c_2$  if  $P(c_1) > P(c_2)$  in any distribution of the posterior credal set.

If no class dominates  $c_1$ , then  $c_1$  is non-dominated.

- If there are more non-dominated classes, CMA returns all of them: the classification is *indeterminate*.
- CMA becomes indeterminate on the instances whose classification would be prior-dependent (and hence fragile) using BMA.



# Practical behavior of CMA

- CMA discriminates between:
  - *easy-to-classify* instances, over which a single class is returned;
  - *hard-to-classify* instances, over which an indeterminate classification is issued.

## Remarks

- CMA, when determinate, returns the same class as BMA, as the credal set includes the flat prior used by BMA;
- CMA will converge to BMA with increasing size of the learning set, as the set of posteriors will converge towards a single posterior;
- the indeterminacy of CMA will hence decrease as more data are available.

# Indicators of performance

## CMA

- *determinacy* (% of determinate classifications);
- *single-accuracy* (% of determ. classification that are accurate);
- *set-accuracy* (% of indeterm. classifications that contain the true class);
- *size of indeterminate output*, i.e., avg. number of classes returned when indeterminate.

# Comparing BMA and CMA

## Indicators

- BMA(CMA D): accuracy of BMA on instances **determinately** classified by CMA.
- BMA(CMA I): accuracy of BMA on instances **indeterminately** classified by CMA.
- If CMA is good at separating easy and hard instances, we should observe  $\text{BMA}(\text{CMA D}) > \text{BMA}(\text{CMA I})$ .

## Results on 31 UCI data sets

- Numerical features are discretized.
- 10 runs of 10-folds cross-validation.
- The reported indicators are averaged over the 31 data sets.

### BMA vs CMA

- BMA (CMA D): 86%
- BMA (CMA I ): 54%

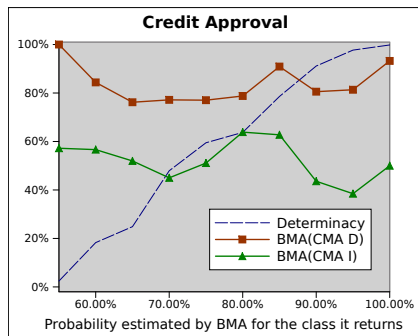
On each data set and setup:  
 $BMA(CMA D) > BMA(CMA I)$

### CMA

- determinacy: 77%
- set-accuracy: 90%
- imprecise output size:  
 $\cong 33\%$  of the classes

- Indeterminate classifications do preserve the reliability of CMA!

# Posterior probabilities vs indeterminate classifications



- Higher posterior probability of BMA leads to higher determinacy of CMA.
- At any level of posterior probability,  $BMA(CMA D) > BMA(CMA I)$ .

# Summary

- CMA extends BMA to imprecise probabilities, to robustly deal with the specification of the prior over the models.
- CMA becomes indeterminate on instances whose classification is prior-dependent, and over which the BMA draws unreliable conclusions.
- Indeterminate classifications preserve the classifier's reliability while conveying sensible information.
- Challenge: develop CMA also for other families of models.