Regularization and Averaging of the Selective Naive Bayes Classifier

Marc Boullé

IJCNN 2006 - Vancouver
Outline

- Selective Naive Bayes
  - principles
  - regularization of variable selection
  - averaging of the models

- Evaluation on 40 datasets

- Results on the Performance Prediction Challenge
Naive Bayes classifier: principle

- The Bayesian classifier is optimal
  - assigns the most probable class given the data
  - but not computable

- Naive Bayes assumption
  - the input variables are independent within each class label
  - "idiot" assumption, but easy to compute

- Reported performances
  - robust and often effective on many real data applications
  - good ranking (but poor estimation) of the class conditional probabilities
An illustrative example: the waveform dataset

The Waveform dataset
- 5000 instances, 21 continuous attributes, 3 classes

Naive Bayes classifier
- univariate class conditional probabilities evaluated using discretization
- two variables removed, 19 variables used by the model

Evaluation on a test set
- ACC: Accuracy
- AUC: Area under the ROC Curve
- ILF: Information Loss Function
Selective Naive Bayes: principles

■ Objective
  ■ remove non-informative or redundant variables

■ Variable selection: standard approach
  ■ selection criterion: accuracy or AUC
  ■ optimization algorithm: forward selection

■ Evaluation on the waveform dataset
Selective Naive Bayes: regularization

- **Objective**
  - control overfitting caused by the variable selection
  - improve the time complexity of the algorithm

- **Regularization** (detailed in the paper)
  - selection criterion: resulting from a MAP approach ($\max (P(\text{Model})P(\text{Data/Model}))$)
  - optimization algorithm: fast forward-backward selection, embedded in multi-start

- **Evaluation on the waveform dataset**

![Graphs showing performance metrics for different models: SNB(MAP), SNB(AUC), SNB(ACC), and NB. The graphs display AUC, ACC, and ILF values for various input features.]
Selective Naive Bayes: averaging with model posterior probabilities

- **Objective**
  - account for model uncertainty in variable selection

- **Model averaging** (detailed in the paper)
  - Bayesian Model Averaging approach: models weight by their posterior probability
  - optimization algorithm: models collected during the MAP optimization

- **Evaluation on the waveform dataset**
Bayesian Model Averaging: understanding why it is not effective

- Using the posterior distribution of the model looks promising
  - 500 best models displayed
  - less than 10 variables used on average

- But this distribution is too sharply peaked
  - 500,000 models evaluated
  - MAP is
    - 40 times more probable than the 3rd model
    - 5,000 times more probable than the 10th model
    - $10^{1033}$ times more probable than the null model
Selective Naive Bayes: averaging with model compression coefficient

- **Objective**
  - better exploit the whole distribution of models

- **New model weighting schema**  (detailed in the paper)
  - use of compression coefficient: ability of the model to compress the class labels given the input data
  - optimization algorithm: same as before
  - model produced: one single Naive Bayes with weighted variables

- **Evaluation on the waveform dataset**
Evaluation on 40 datasets
Summary of the datasets

- 40 datasets
  - 30 datasets from the UCI repository
  - 10 datasets from the Feature Selection and Performance Prediction Challenges

- Large diversity
  - 100 to 50,000 instances
  - 4 to 100,000 input variables
  - Continuous or categorical
  - 2 to 28 class labels
Evaluation protocol

- Evaluated Naive Bayes methods
  - no variable selection
    - NB(EF): discretization using 10 bins Equal Frequency
    - NB: discretization and value grouping using MODL optimal methods (Boullé, 2006)
  - variable selection
    - SNB(ACC): optimization of accuracy
    - SNB(AUC): optimization of AUC
    - SNB(MAP): MAP regularization
  - variable selection and model averaging
    - SNB(MA): weights using model posterior probabilities
    - SNB(CMA): weights using model compression coefficients

- Evaluated criteria, using stratified 10-cross validation
  - ACC: evaluation of the prediction
  - AUC: evaluation of the ranking of the class conditional probabilities
  - ILF: evaluation of the class conditional probabilities
Results on the UCI datasets

![Graphs showing performance metrics for different algorithms on UCI datasets.]
Results on the Challenge datasets
Evaluation on the Performance Prediction Challenge
Objective of the Challenge

- Bi-criterion evaluation
  - predict the class labels
  - predict the prediction performance

- Five datasets
  - 3000 to 13000 training instances
  - 50 to 17000 continuous input variables
  - 2 class values
Method used in the challenge (1/2)

- Same method for all datasets: SNB(CMA)
  - input variables discretized using the MODL method (Boullé, 2006)
  - variable selection, with MAP regularization
  - model averaging with model compression coefficients

- Post-optimization for the challenge
  - modify class priors to reach better BER criterion

- Training
  - using all the available data

- Performance prediction
  - using stratified 10-cross validation
Method used in the challenge (2/2)

Feature construction
- to evaluate the computational scalability
- to evaluate the statistical scalability (overfitting behavior)
- to leverage the Naive Bayes assumption

Four feature construction schema experimented
- SNB(CMA): no feature construction
- SNB(CMA) 10k F(2D): 10 000 features constructed
  - sums of two randomly chosen input variables
- SNB(CMA) 100k F(2D): 100 000 features constructed
  - sums of two randomly chosen input variables
- SNB(CMA) 10k F(3D): 10 000 features constructed
  - sums of three randomly chosen input variables
Overall results

- Overall challenge ranking
  - Average rank: 7th
  - 1st on two datasets

- By criterion:
  - Test BER: 12th
  - Test AUC: 3rd
  - Test Guess Error: 6th

![Overall results graph](image-url)
Ada: Marketing

- SNB(CMA) method ranked first

![Graph showing Ada results with different error rates and BER values.](image-url)
Gina: Digit recognition

- Poor BER performance, caused by the Naive Bayes assumption
- Feature construction helps a lot: up to 6% improvement
Hiva: Drug discovery

- Poor BER performance

![Graph showing Hiva results with different categories and test BER values]
Nova: Text classification

- Poor performance

![Diagram showing Nova results with various test BER and guess error categories. The diagram includes points for different entry types and their corresponding error measurements.](image-url)
Sylva: Ecology

- SNB(CMA) method ranked first
Training time

- Training time: linear in the number of values \((\text{instances} \times \text{variables})\)
  - evaluated up to 1 billion values
Analysis of the results

- **Performance**
  - very efficient when subsets of variables are consistent with the naive assumption
  - sometimes competitive with much more sophisticated methods
  - but limited by the "idiot" assumption

- **Computational scalability**
  - highly scalable
  - linear in the number of values (instances*variables)

- **Statistical scalability (overfitting behavior)**
  - adding huge numbers of features never degrades the performances

- **Feature construction to leverage the Naive Bayes assumption**
  - feature construction sometimes brings significant improvements (Digits)
  - random feature construction is not very efficient
Conclusion
Conclusion

- **SNB(CMA) method principles**
  - exploit the Naive Bayes assumption
  - preprocessing of variables using the MODL method
  - regularization of the variable selection
  - efficient optimization heuristic
  - model averaging using compression weights

- **SNB(CMA) performance**
  - highly scalable and resistant to overfitting
  - significantly improved estimation of the class conditional probabilities
  - main limitation: the Naive Bayes assumption

- **Future work**
  - apply the same approach on more sophisticated models
The software is now available

Data preparation & scoring tool for supervised learning

• Download: