Combining Naive Bayes and Decision Tables

Mark Hall[1] and Eibe Frank[2]

[1] *Pentaho Corporation, 5950 Hazeltine National Drive, Suite 340, Orlando, FL, USA*
[2] *Department of Computer Science, University of Waikato, New Zealand*

Introduction

- Our combined model is a semi-naive Bayesian ranking method that combines naive Bayes with decision tables.
- Is a simple Bayesian network in which the decision table represents a conditional probability table.
- Can be viewed as a restricted version of Pazzani's semi-naive Bayesian model that finds one, rather than multiple, groups of dependent attributes.
- Has lower computational complexity that Pazanni's method.
- Search and evaluation is based on AUC.
- Empirical results show that the ranker resulting from our combined model, compared to either component technique, frequently significantly increases AUC.

Decision Tables (DT)

- Store the input data in condensed form based on a selected set of attributes.
- Is essentially a lookup table when making predictions.
- Each entry in the table is associated with class probability estimates based on observed frequencies.
- Cross-validation is used to choose a set of discriminative attributes for the table.
- Cross-validation is efficient as the *structure* of the table does not change when adding or deleting instances.
- In our experiments we used forward selection (guided by AUC) to select attributes for stand-alone decision tables.
- Numeric attributes were discretized using MDL-based discretization.

Naive Bayes (NB)

- Simple and fast learner.
- Computes the posterior probability of a class using Bayes theorem.
- Probabilities for attribute values conditioned on the class are computed using frequency counts from the training data.
- Efficient under cross-validation as frequency counts can be updated in constant time.
- Numeric attributes were discretized using MDL-based discretization.
- In our experiments we used standard naive Bayes and a version that uses forward selection, guided by AUC, to select attributes (NB$_{AS}$).

Combined Model (DTNB)

- Learning the combined model is similar to learning a decision table.
- At each step in the search:
  1. Split the attributes into two disjoint subsets: one for the decision table, the other for naive Bayes.
  2. Evaluate the merit of the combined model based on the split.
- We use a forward selection search where:
  - At each step, selected attributes are modeled by naive Bayes and the remainder by the decision table.
  - Initially, all attributes are modeled by the decision table.
  - Leave-one-out cross-validated AUC is used to evaluate the quality of a split based on the probability estimates generated by the combined model.

Combined Model (DTNB)

- Combining class probability estimates from the decision table and naive Bayes:

$$Q(y|X) = \alpha \times Q_{DT}(y|X^{\top}) \times Q_{NB}(y|X^{\perp})/Q(y),$$

where:
  - $Q_{DT}(y|X^{\top})$ and $Q_{NB}(y|X^{\perp})$ are the class probability estimates obtained from the DT and NB respectively.
  - α is a normalization constant.
  - $Q(y)$ is the prior probability of the class.
- Probabilities are estimated using Laplace-corrected observed counts.
- We also consider a variant of the combined model that includes attribute selection (DTNB$_{AS}$).

Data Sets

Dataset	Instances	Attributes	Classes
anneal	898	38	5
autos	205	25	6
balance-s	625	4	3
breast-c	286	9	2
breast-w	699	9	2
credit-a	690	15	2
credit-g	1000	20	2
diabetes	768	8	2
ecoli	336	7	8
glass	214	9	6
heart-c	303	13	2
heart-h	294	13	2
heart-s	270	13	2
hepatitis	155	19	2
horse-c	368	22	2
hypothyroid	3772	29	4
ionosphere	351	34	2
iris	150	4	3
kr-vs-kp	3196	36	2
labor	57	16	2
lymphography	148	18	4
mushroom	8124	22	2
optdigits	5620	64	10
pendigits	10992	16	10
primary-t	339	17	21
segment	2310	19	7
sick	3772	29	2
sonar	208	60	2
soybean	683	35	19
splice	3190	61	3
vehicle	846	18	4
vote	435	16	2
vowel	990	13	11
waveform	5000	40	3
zoo	101	16	7

Results: Mean AUC w/o attribute selection

Dataset	DTNB	NB	DT
anneal	0.9970±0.0080	0.9773±0.0138 •	0.9986±0.0037
autos	0.8887±0.0772	0.8613±0.0818	0.9233±0.0569
balance-s	0.9666±0.0192	0.9035±0.0374 •	0.9129±0.0370 •
breast-c	0.6669±0.1090	0.6901±0.1060	0.6432±0.1149
breast-w	0.9922±0.0075	0.9920±0.0076	0.9845±0.0118 •
credit-a	0.9266±0.0318	0.9253±0.0310	0.9199±0.0342
credit-g	0.7554±0.0438	0.7812±0.0522 ○	0.7006±0.0588 •
diabetes	0.8037±0.0573	0.8053±0.0569	0.7971±0.0578
ecoli	0.9868±0.0158	0.9865±0.0150	0.9819±0.0176
glass	0.7485±0.1100	0.7487±0.1036	0.7481±0.1076
heart-c	0.9083±0.0462	0.9109±0.0478	0.8656±0.0524 •
heart-h	0.9206±0.0474	0.9205±0.0487	0.8900±0.0583 •
heart-s	0.8861±0.0612	0.8959±0.0618	0.8777±0.0714
hepatitis	0.8984±0.1063	0.9080±0.1004	0.7767±0.1331 •
horse-c	0.8713±0.0752	0.8365±0.0820	0.8721±0.0478
hypothyroid	0.9950±0.0050	0.9945±0.0035	0.9979±0.0024
ionosphere	0.9533±0.0313	0.9512±0.0302	0.9036±0.0522 •
iris	1.0000±0.0000	1.0000±0.0000	1.0000±0.0000
kr-vs-kp	0.9926±0.0029	0.9525±0.0104 •	0.9946±0.0036 ○
labor	0.9600±0.0762	0.9608±0.0750	0.8633±0.1336
lymphography	0.9202±0.0615	0.9208±0.0584	0.8881±0.0768
mushroom	1.0000±0.0000	0.9981±0.0007 •	1.0000±0.0000
optdigits	0.9909±0.0060	0.9838±0.0066 •	0.9629±0.0132 •
pendigits	0.9919±0.0022	0.9869±0.0028 •	0.9891±0.0038 •
primary-t	0.8777±0.0590	0.8967±0.0503 ○	0.8677±0.0609
segment	0.9992±0.0013	0.9986±0.0020	0.9977±0.0028
sick	0.9560±0.0204	0.9555±0.0199	0.9500±0.0244
sonar	0.8719±0.0725	0.8874±0.0581	0.8255±0.0883
soybean	0.9902±0.0127	0.9656±0.0280 •	0.9649±0.0471
splice	0.9831±0.0048	0.9771±0.0052 •	0.9655±0.0087 •
vehicle	0.9762±0.0144	0.9388±0.0249 •	0.9716±0.0144
vote	0.9886±0.0132	0.9745±0.0191 •	0.9856±0.0129
vowel	0.9967±0.0052	0.9914±0.0107	0.9923±0.0113
waveform	0.9485±0.0100	0.9422±0.0102 •	0.8938±0.0151 •
zoo	1.0000±0.0000	1.0000±0.0000	1.0000±0.0000

•, ○ statistically significant improvement or degradation for DTNB

Results: Mean AUC with attribute selection

Dataset	DTNB$_{AS}$	NB$_{AS}$	DT
anneal	0.9983±0.0075	0.9882±0.0163 •	0.9986±0.0037
autos	0.8934±0.0751	0.8724±0.0848	0.9233±0.0569
balance-s	0.9666±0.0192	0.9669±0.0192	0.9129±0.0370 •
breast-c	0.6615±0.1095	0.6718±0.1083	0.6432±0.1149
breast-w	0.9920±0.0078	0.9910±0.0086	0.9845±0.0118 •
credit-a	0.9298±0.0332	0.9287±0.0318	0.9199±0.0342
credit-g	0.7577±0.0462	0.7788±0.0512 ○	0.7006±0.0588 •
diabetes	0.8024±0.0589	0.8049±0.0570	0.7971±0.0578
ecoli	0.9870±0.0153	0.9871±0.0152	0.9819±0.0176
glass	0.7487±0.1100	0.7493±0.1087	0.7481±0.1076
heart-c	0.9105±0.0468	0.9094±0.0474	0.8656±0.0524 •
heart-h	0.9233±0.0468	0.9197±0.0518	0.8900±0.0583 •
heart-s	0.8831±0.0564	0.8979±0.0633	0.8777±0.0714
hepatitis	0.8960±0.1089	0.8930±0.1045	0.7767±0.1331 •
horse-c	0.8715±0.0757	0.8740±0.0786	0.8721±0.0478
hypothyroid	0.9956±0.0038	0.9968±0.0026	0.9979±0.0024
ionosphere	0.9568±0.0282	0.9596±0.0239	0.9036±0.0522 •
iris	1.0000±0.0000	1.0000±0.0000	1.0000±0.0000
kr-vs-kp	0.9952±0.0024	0.9870±0.0046 •	0.9946±0.0036
labor	0.9575±0.0920	0.9717±0.0822	0.8633±0.1336
lymphography	0.9300±0.0586	0.9185±0.0628	0.8881±0.0768
mushroom	1.0000±0.0000	0.9999±0.0001 •	1.0000±0.0000
optdigits	0.9909±0.0059	0.9927±0.0046	0.9629±0.0132 •
pendigits	0.9936±0.0018	0.9892±0.0026 •	0.9891±0.0038 •
primary-t	0.8770±0.0609	0.8848±0.0567	0.8677±0.0609
segment	0.9994±0.0012	0.9987±0.0019	0.9977±0.0028
sick	0.9544±0.0205	0.9563±0.0196	0.9500±0.0244
sonar	0.8699±0.0703	0.8862±0.0703	0.8255±0.0883
soybean	0.9900±0.0115	0.9930±0.0116	0.9649±0.0471
splice	0.9841±0.0044	0.9823±0.0050 •	0.9655±0.0087 •
vehicle	0.9807±0.0150	0.9680±0.0175 •	0.9716±0.0144
vote	0.9905±0.0096	0.9906±0.0080	0.9856±0.0129
vowel	0.9970±0.0051	0.9941±0.0066	0.9923±0.0113
waveform	0.9479±0.0099	0.9455±0.0098 •	0.8938±0.0151 •
zoo	1.0000±0.0000	1.0000±0.0000	1.0000±0.0000

•, ○ statistically significant improvement or degradation for DTNB$_{AS}$

Experiments

- 35 UCI data sets.
- Multi-class data sets were converted to two-class data sets by merging all classes except the largest one.
- 50 runs of repeated holdout (66% training).
- Report mean AUC and standard deviation.
- Identical runs were used for each algorithm.
- Statistical significance computed from the corrected resampled t-test at the 5% level.

Conclusions

- The combined model (DTNB) is a simple and efficient semi-naive Bayesian ranking algorithm.
- Input attributes are split into two groups: one group assigns class probabilities based on naive Bayes, the other group based on a decision table, and the resulting probability estimates are combined.
- Empirical results show that:
  1. DTNB performs well compared to stand-alone naive Bayes and decision tables. 11 significant wins against both, with two and one significant loss respectively.
  2. There are five cases where DTNB is significantly better than both.
  3. When attribute selection is applied to both DTNB and naive Bayes there are seven significant wins for DTNB and only one significant loss.
  4. Applying attribute selection to naive Bayes renders its computational equal to DTNB (quadratic in the number of attributes).
  5. Compared to standard decision tables, which have built-in attribute selection, DTNB achieves 11 wins and no significant losses.