Data Mining 2012
Bayesian Network Classifiers

Ad Feelders

Universiteit Utrecht

October 25, 2012
N. Friedman, D. Geiger and M. Goldszmidt
Bayesian Network Classifiers
Machine Learning, 29, pp. 131-163 (1997)
(except section 6)
Bayesian Networks are models of the *joint distribution* of a collection of random variables. The joint distribution is simplified by introducing independence assumptions.

In many applications we are in fact interested in the conditional distribution of one variable (the class variable) given the other variables (attributes).

Can we use Bayesian Networks as classifiers?
The Naive Bayes Classifier

This Bayesian Network is equivalent to its undirected version (why?):

Attributes are independent given the class label.
BN factorisation:

\[ P(X) = \prod_{i=1}^{k} P(X_i \mid X_{pa(i)}) , \]

So factorisation corresponding to NB classifier is:

\[ P(C, A_1, \ldots, A_k) = P(C)P(A_1 \mid C) \cdots P(A_k \mid C) \]
Naive Bayes assumption

Via Bayes rule we have

\[ P(C = i | A) = \frac{P(A_1, A_2, \ldots, A_k, C = i)}{P(A_1, A_2, \ldots, A_k)} \]  

(product rule)

\[ = \frac{P(A_1, A_2, \ldots, A_k | C = i)P(C = i)}{\sum_{j=1}^{c} P(A_1, A_2, \ldots, A_k | C = j)P(C = j)} \]  

(product rule and sum rule)

\[ = \frac{P(A_1 | C = i)P(A_2 | C = i) \cdots P(A_k | C = i)P(C = i)}{\sum_{j=1}^{c} P(A_1 | C = j)P(A_2 | C = j) \cdots P(A_k | C = j)P(C = j)} \]  

(NB factorisation)
Why Naive Bayes is competitive

The conditional independence assumption is often clearly inappropriate, yet the *predictive accuracy* of Naive Bayes is competitive with more complex classifiers. How come?

- Probability estimates of Naive Bayes may be way off, but this does not necessarily result in wrong classification!
- Naive Bayes has only few parameters compared to more complex models, so it can estimate parameters more reliably.
Naive Bayes: Example

\[ P(C = 0) = 0.4, \quad P(C = 1) = 0.6 \]

\[
\begin{array}{c|cc|c}
C = 0 & A_2 & P(A_1) \\
A_1 & 0 & 1 & \\
0 & 0.2 & 0.1 & 0.3 \\
1 & 0.1 & 0.6 & 0.7 \\
P(A_2) & 0.3 & 0.7 & 1 \\
\end{array}
\quad
\begin{array}{c|cc|c}
C = 1 & A_2 & P(A_1) \\
A_1 & 0 & 1 & \\
0 & 0.5 & 0.2 & 0.7 \\
1 & 0.1 & 0.2 & 0.3 \\
P(A_2) & 0.6 & 0.4 & 1 \\
\end{array}
\]

We have that

\[
P(C = 1|A_1 = 0, A_2 = 0) = \frac{0.5 \times 0.6}{0.5 \times 0.6 + 0.2 \times 0.4} = 0.79
\]

According to naive Bayes

\[
P(C = 1|A_1 = 0, A_2 = 0) = \frac{0.7 \times 0.6 \times 0.6}{0.7 \times 0.6 \times 0.6 + 0.3 \times 0.3 \times 0.4} = 0.88
\]

Naive Bayes assigns to the right class.
What about this model?

BN factorisation:

\[ P(X) = \prod_{i=1}^{k} P(X_i | X_{pa(i)}) , \]

So factorisation is:

\[ P(C, A_1, \ldots, A_k) = P(C|A_1, \ldots, A_k) P(A_1) \cdots P(A_k) \]
Markov Blanket of $C$: Moral Graph


Local Markov property: $C \perp \perp \text{rest} \mid \text{boundary}(C)$
Bayesian Networks as Classifiers

Loglikelihood under model $M$ is

$$\mathcal{L}(M|D) = \sum_{j=1}^{n} \log P_M(X(j))$$

where $X(j) = (A_1^{(j)}, A_2^{(j)}, \ldots, A_k^{(j)}, C^{(j)})$.

We can rewrite this as

$$\mathcal{L}(M|D) = \sum_{j=1}^{n} \log P_M(C^{(j)}|A^{(j)}) + \sum_{j=1}^{n} \log P_M(A^{(j)})$$

If there are many attributes, the second term will dominate the loglikelihood score.

But we are not interested in modeling the distribution of the attributes!
Bayesian Networks as Classifiers

-\log P(x)

P(x)

Ad Feelders (Universiteit Utrecht)

Data Mining

October 25, 2012
Naive Bayes vs. Unrestricted BN

To confirm this hypothesis, we conducted an experiment comparing the classification accuracy of Bayesian networks learned using the MDL score (i.e., classifiers based on unrestricted networks) to that of the naive Bayesian classifier. We ran this experiment on...
Use Conditional Log-likelihood?

*Discriminative vs. Generative* learning.

Conditional loglikelihood function:

\[
\mathcal{L}(M|D) = \sum_{j=1}^{n} \log P_M(C^{(j)}|A_1^{(j)}, \ldots, A_k^{(j)})
\]

No closed form solution for ML estimates.

Remark: can be done via Logistic Regression for models with *perfect* graphs (Naive Bayes, TAN’s).
NB and Logistic Regression

The logistic regression assumption is

\[
\log \left\{ \frac{P(C = 1 | A)}{P(C = 0 | A)} \right\} = \alpha + \sum_{i=1}^{k} \beta_i A_i,
\]

that is, the log odds is a linear function of the attributes.

Under the naive Bayes assumption, this is exactly true.

Assign to class 1 if \( \alpha + \sum_{i=1}^{k} \beta_i A_i > 0 \) and to class 0 otherwise.

Logistic regression maximizes conditional likelihood under this assumption (it is a so-called discriminative model).

There is no closed form solution for the maximum likelihood estimates of \( \alpha \) and \( \beta_i \), but the loglikelihood function is globally concave (unique global optimum).
Proof (for binary attributes $A_i$)

Under the naive Bayes assumption we have:

$$\frac{P(C = 1|a)}{P(C = 0|a)} = \frac{P(a_1|C = 1) \cdots P(a_k|C = 1)P(C = 1)}{P(a_1|C = 0) \cdots P(a_k|C = 0)P(C = 0)}$$

$$= \prod_{i=1}^{k} \left\{ \left( \frac{P(a_i = 1|C = 1)}{P(a_i = 1|C = 0)} \right)^{a_i} \left( \frac{P(a_i = 0|C = 1)}{P(a_i = 0|C = 0)} \right)^{1-a_i} \right\}$$

$$\times \frac{P(C = 1)}{P(C = 0)}$$

Taking the log we get

$$\log \left( \frac{P(C = 1|a)}{P(C = 0|a)} \right) = \sum_{i=1}^{k} \left\{ a_i \log \left( \frac{P(a_i = 1|C = 1)}{P(a_i = 1|C = 0)} \right) + (1-a_i) \log \left( \frac{P(a_i = 0|C = 1)}{P(a_i = 0|C = 0)} \right) \right\} + \log \left( \frac{P(C = 1)}{P(C = 0)} \right)$$
Proof (continued)

Expand and collect terms.

\[
\log \left( \frac{P(C = 1|a)}{P(C = 0|a)} \right) = \sum_{i=1}^{k} a_i \left\{ \log \frac{P(a_i = 1|C = 1)}{P(a_i = 1|C = 0)} \frac{P(a_i = 0|C = 0)}{P(a_i = 0|C = 1)} \right\} + \sum_{i=1}^{k} \left\{ \log \frac{P(a_i = 0|C = 1)}{P(a_i = 0|C = 0)} \right\} + \log \frac{P(C = 1)}{P(C = 0)}
\]

which is a linear function of \(a\).
Example

Suppose $P(C = 1) = 0.6, P(a_1 = 1|C = 1) = 0.8, P(a_1 = 1|C = 0) = 0.5, P(a_2 = 1|C = 1) = 0.6, P(a_2 = 1|C = 0) = 0.3.$

Then

$$\log \left( \frac{P(C = 1|a_1, a_2)}{P(C = 0|a_1, a_2)} \right) = 1.386a_1 + 1.253a_2 - 1.476 + 0.405$$

$$= -1.071 + 1.386a_1 + 1.253a_2$$

Classify a point with $a_1 = 1$ and $a_2 = 0$:

$$\log \left( \frac{P(C = 1|1, 0)}{P(C = 0|1, 0)} \right) = -1.071 + 1.386 \times 1 + 1.253 \times 0 = 0.315$$

Decision rule: assign to class 1 if

$$\alpha + \sum_{i=1}^{k} \beta_i A_i > 0$$

and to class 0 otherwise. Linear decision boundary.
A_2 = 0.855 - 1.106A_1
Relax strong assumptions of NB

- Conditional independence assumption of NB is often incorrect, and could lead to suboptimal classification performance.
- Relax this assumption by allowing (restricted) dependencies between attributes.
- This may produce more accurate probability estimates, possibly leading to better classification performance.
- This is not guaranteed, because the more complex model may be overfitting.
Tree Structured BN

For tree structured Bayesian Networks there is an algorithm, due to Chow and Liu, that produces the optimal structure in polynomial time.

This algorithm is guaranteed to produce the tree structure that maximizes the loglikelihood score.

Tree structure: each node (except the root of the tree) has exactly one parent.

Why no penalty for complexity?
Mutual Information

Measure of association between (discrete) random variables $X$ and $Y$:

$$I_P(X; Y) = \sum_{x,y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)}$$

“Distance” (Kullback-Leibler divergence) between joint distribution of $X$ and $Y$, and their joint distribution under the independence assumption.

If $X$ and $Y$ are independent, their mutual information is zero, otherwise it is some positive quantity.
Algorithm Construct-Tree of Chow and Liu

- Compute $I_{\hat{P}_D}(X_i; X_j)$ between each pair of variables. \(O(nk^2)\)
- Build complete undirected graph with weights $I_{\hat{P}_D}(X_i; X_j)$
- Build a maximum weighted spanning tree \(O(k^2 \log k)\)
- Choose root, and let all edges point away from it
### Example Data Set

<table>
<thead>
<tr>
<th>obs</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>
\[
I_{\hat{P}_D}(X_1; X_4) = \sum_{x_1, x_4} \hat{P}_D(x_1, x_4) \log \frac{\hat{P}_D(x_1, x_4)}{\hat{P}_D(x_1)\hat{P}_D(x_4)}
\]

\[
= 0.4 \log \frac{0.4}{(0.5)(0.4)} + 0.1 \log \frac{0.1}{(0.5)(0.2)} + 0 \log \frac{0}{(0.5)(0.4)} + 0 \log \frac{0}{(0.5)(0.4)}
\]

\[
+ 0.1 \log \frac{0.1}{(0.5)(0.2)} + 0.4 \log \frac{0.4}{(0.5)(0.4)}
\]

\[
= 0.55
\]
Build Graph with Weights

```
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.032</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.032</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.55</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0.032</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th></th>
<th></th>
<th>4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.032</td>
<td></td>
<td></td>
<td>0.032</td>
</tr>
</tbody>
</table>
```

Ad Feelders (Universiteit Utrecht)
Maximum weighted spanning tree
Choose root node
Algorithm to construct TAN

- Compute $I_{\hat{P}_D}(A_i; A_j|C)$ between each pair of attributes, where

$$I_P(A_i; A_j|C) = \sum_{a_i,a_j,c} P(a_i, a_j, c) \log \frac{P(a_i, a_j|c)}{P(a_i|c)P(a_j|c)}$$

is the conditional mutual information between $A_i$ and $A_j$ given $C$.
- Build complete undirected graph with weights $I_{\hat{P}_D}(A_i; A_j|C)$
- Build a maximum weighted spanning tree
- Choose root, and let all edges point away from it
- Construct a TAN by adding $C$ and an arc from $C$ to all attributes.

This algorithm is guaranteed to produce the TAN structure with optimal loglikelihood score.
Figure 3. A TAN model learned for the data set “pima.” The dashed lines are those edges required by the naive Bayesian classifier. The solid lines are correlation edges between attributes.

These approaches, however, remove the strong assumptions of independence in naive Bayes by finding correlations among attributes that are warranted by the training data.

4.1. Augmented naive Bayesian networks as classifiers

We argued above that the performance of a Bayesian network as a classifier may improve if the learning procedure takes into account the special status of the class variable. An easy way to ensure this is to bias the structure of the network, as in the naive Bayesian classifier, such that there is an edge from the class variable to each attribute. This ensures that, in the learned network, the probability $P(C | A_1, ..., A_n)$ will take all attributes into account. In order to improve the performance of a classifier based on this bias, we propose to augment the naive Bayes structure with edges among the attributes, when needed, thus dispensing with its strong assumptions about independence. We call these structures augmented naive Bayesian networks and these edges augmenting edges.

In an augmented structure, an edge from $A_i$ to $A_j$ implies that the influence of $A_i$ on the assessment of the class variable also depends on the value of $A_j$. For example, in Figure 3, the influence of the attribute “Glucose” on the class $C$ depends on the value of “Insulin,” while in the naive Bayesian classifier the influence of each attribute on the class variable is independent of other attributes. These edges affect the classification process in that a value of “Glucose” that is typically surprising (i.e., $P(g | c)$ is low) may be unsurprising if the value of its correlated attribute, “Insulin,” is also unlikely (i.e., $P(g | c, i)$ is high). In this situation, the naive Bayesian classifier will overpenalize the probability of the class variable by considering two unlikely observations, while the augmented network of Figure 3 will not.

Adding the best set of augmenting edges is an intractable problem, since it is equivalent to learning the best Bayesian network among those in which $C$ is a root. Thus, even if we could improve the performance of a naive Bayes classifier in this way, the computational effort required may not be worthwhile. However, by imposing acceptable restrictions on the form of the allowed interactions, we can actually learn the optimal set of augmenting edges in polynomial time.
Interpretation of TAN’s

Just like NB models, TAN’s are equivalent to their undirected counterparts (why?).

Since there is an edge between Pregnant and Age, the influence of Age on the class label depends on (is different for different values of) Pregnant (and vice versa).
Smoothing by adding “prior counts”

Sometimes we have few observations to estimate (conditional) probabilities.

Add “prior counts” to “smooth” the estimates.

\[
\hat{p}^s(x_i \mid x_{pa(i)}) = \frac{n(x_{pa(i)})\hat{p}(x_i \mid x_{pa(i)}) + m(x_{pa(i)})p^0(x_i \mid x_{pa(i)})}{n(x_{pa(i)}) + m(x_{pa(i)})}
\]

where \(m(x_{pa(i)})\) is the prior precision, \(\hat{p}^s(x_i \mid x_{pa(i)})\) is the smoothed estimate, and \(p^0(x_i \mid x_{pa(i)})\) is our prior estimate of \(p(x_i \mid x_{pa(i)})\).

Common to take \(m(x_{pa(i)})\) to be the same for all parent configurations.

Weighted average of ML estimate and prior estimate.
ML estimates

For example

\[ \hat{p}_1(1) = \frac{n(x_1 = 1)}{n} = \frac{5}{10} = \frac{1}{2} \]

and

\[ \hat{p}_{3|1,2}(1|1,2) = \frac{n(x_1 = 1, x_2 = 2, x_3 = 1)}{n(x_1 = 1, x_2 = 2)} = \frac{0}{2} = 0 \]
Suppose we set $m = 2$, and $p^0(x_i \mid x_{pa(i)}) = \hat{p}(x_i)$.

Then we get

$$\hat{p}_1^s(1) = \frac{10 \times 0.5 + 2 \times 0.5}{10 + 2} = \frac{1}{2}$$

and

$$\hat{p}_{3|1,2}^s(1|1,2) = \frac{2 \times 0 + 2 \times 0.4}{2 + 2} = 0.2$$
Bayesian Multinets

Build structure on attributes for each class separately and use

\[ P_M(C = i, A_1, \ldots, A_k) = P(C = i)P_M(A_1, \ldots, A_k | C = i) \]

\[ = P(C = i)P_{M_i}(A_1, \ldots, A_k) \quad i = 1, \ldots, c \]

Using trees for class-conditional structures

- Split \( D \) into \( c \) partitions, \( D_1, D_2, \ldots, D_c \) where \( c \) is the number of distinct values of class label \( C \). \( D_i \) contains all records in \( D \) with \( C = i \).
- Set \( P(C = i) = \hat{P}_D(C = i) \) for \( i = 1, \ldots, c \).
- Apply Construct-Tree on \( D_i \) to construct \( M_i \).
The following classifiers were compared:

- NB: Naive Bayes Classifier
- SNB: Naive Bayes with attribute selection
- BN: Unrestricted Bayesian Network (Markov Blanket of Class)
- C4.5: Classification Tree
- And also

<table>
<thead>
<tr>
<th>structure</th>
<th>same</th>
<th>different</th>
</tr>
</thead>
<tbody>
<tr>
<td>tree</td>
<td>TAN</td>
<td>CL</td>
</tr>
<tr>
<td>dag</td>
<td>ANB</td>
<td>MN</td>
</tr>
</tbody>
</table>

Superscript $s$ indicates smoothing of parameter estimates.
Table 1. Description of data sets used in the experiments.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Attributes</th>
<th># Classes</th>
<th># Instances</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>australian</td>
<td>14</td>
<td>2</td>
<td>690</td>
<td>CV-5</td>
</tr>
<tr>
<td>2</td>
<td>breast</td>
<td>10</td>
<td>2</td>
<td>683</td>
<td>CV-5</td>
</tr>
<tr>
<td>3</td>
<td>chess</td>
<td>36</td>
<td>2</td>
<td>2130</td>
<td>1066</td>
</tr>
<tr>
<td>4</td>
<td>cleve</td>
<td>13</td>
<td>2</td>
<td>296</td>
<td>CV-5</td>
</tr>
<tr>
<td>5</td>
<td>corral</td>
<td>6</td>
<td>2</td>
<td>128</td>
<td>CV-5</td>
</tr>
<tr>
<td>6</td>
<td>crx</td>
<td>15</td>
<td>2</td>
<td>653</td>
<td>CV-5</td>
</tr>
<tr>
<td>7</td>
<td>diabetes</td>
<td>8</td>
<td>2</td>
<td>768</td>
<td>CV-5</td>
</tr>
<tr>
<td>8</td>
<td>flare</td>
<td>10</td>
<td>2</td>
<td>1066</td>
<td>CV-5</td>
</tr>
<tr>
<td>9</td>
<td>german</td>
<td>20</td>
<td>2</td>
<td>1000</td>
<td>CV-5</td>
</tr>
<tr>
<td>10</td>
<td>glass</td>
<td>9</td>
<td>7</td>
<td>214</td>
<td>CV-5</td>
</tr>
<tr>
<td>11</td>
<td>glass2</td>
<td>9</td>
<td>2</td>
<td>163</td>
<td>CV-5</td>
</tr>
<tr>
<td>12</td>
<td>heart</td>
<td>13</td>
<td>2</td>
<td>270</td>
<td>CV-5</td>
</tr>
<tr>
<td>13</td>
<td>hepatitis</td>
<td>19</td>
<td>2</td>
<td>80</td>
<td>CV-5</td>
</tr>
<tr>
<td>14</td>
<td>iris</td>
<td>4</td>
<td>3</td>
<td>150</td>
<td>CV-5</td>
</tr>
<tr>
<td>15</td>
<td>letter</td>
<td>16</td>
<td>26</td>
<td>15000</td>
<td>5000</td>
</tr>
<tr>
<td>16</td>
<td>lymphography</td>
<td>18</td>
<td>4</td>
<td>148</td>
<td>CV-5</td>
</tr>
<tr>
<td>17</td>
<td>mofn-3-7-10</td>
<td>10</td>
<td>2</td>
<td>300</td>
<td>1024</td>
</tr>
<tr>
<td>18</td>
<td>pima</td>
<td>8</td>
<td>2</td>
<td>768</td>
<td>CV-5</td>
</tr>
<tr>
<td>19</td>
<td>satimage</td>
<td>36</td>
<td>6</td>
<td>4435</td>
<td>2000</td>
</tr>
<tr>
<td>20</td>
<td>segment</td>
<td>19</td>
<td>7</td>
<td>1540</td>
<td>770</td>
</tr>
<tr>
<td>21</td>
<td>shuttle-small</td>
<td>9</td>
<td>7</td>
<td>3866</td>
<td>1934</td>
</tr>
<tr>
<td>22</td>
<td>soybean-large</td>
<td>35</td>
<td>19</td>
<td>562</td>
<td>CV-5</td>
</tr>
<tr>
<td>23</td>
<td>vehicle</td>
<td>18</td>
<td>4</td>
<td>846</td>
<td>CV-5</td>
</tr>
<tr>
<td>24</td>
<td>vote</td>
<td>16</td>
<td>2</td>
<td>435</td>
<td>CV-5</td>
</tr>
<tr>
<td>25</td>
<td>waveform-21</td>
<td>21</td>
<td>3</td>
<td>300</td>
<td>4700</td>
</tr>
</tbody>
</table>
Table 2. Experimental results of the primary approaches discussed in this paper.

<table>
<thead>
<tr>
<th>Data set</th>
<th>NB</th>
<th>BN</th>
<th>TAN$^s$</th>
<th>CL$^s$</th>
<th>C4.5</th>
<th>SNB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 australian</td>
<td>86.23±1.10</td>
<td>86.23±1.76</td>
<td>84.20±1.24</td>
<td>85.07±1.31</td>
<td>85.65±1.82</td>
<td>86.67±1.81</td>
</tr>
<tr>
<td>2 breast</td>
<td>97.36±0.50</td>
<td>96.92±0.63</td>
<td>96.92±0.67</td>
<td>97.07±0.66</td>
<td>94.73±0.59</td>
<td>96.19±0.63</td>
</tr>
<tr>
<td>3 chess</td>
<td>87.15±1.03</td>
<td>95.59±0.63</td>
<td>92.31±0.82</td>
<td>92.40±0.81</td>
<td>99.53±0.21</td>
<td>94.28±0.71</td>
</tr>
<tr>
<td>4 cleve</td>
<td>82.76±1.27</td>
<td>81.39±1.82</td>
<td>81.76±0.33</td>
<td>80.73±1.40</td>
<td>73.31±0.63</td>
<td>78.06±2.41</td>
</tr>
<tr>
<td>5 corral</td>
<td>85.88±3.25</td>
<td>97.60±2.40</td>
<td>96.06±2.51</td>
<td>99.23±0.77</td>
<td>97.69±2.31</td>
<td>83.57±3.15</td>
</tr>
<tr>
<td>6 crx</td>
<td>86.22±1.14</td>
<td>85.60±0.17</td>
<td>85.76±1.16</td>
<td>86.22±1.14</td>
<td>86.22±0.58</td>
<td>85.92±1.08</td>
</tr>
<tr>
<td>7 diabetes</td>
<td>74.48±0.89</td>
<td>75.39±0.29</td>
<td>75.52±1.11</td>
<td>74.74±1.19</td>
<td>76.04±0.85</td>
<td>76.04±0.83</td>
</tr>
<tr>
<td>8 flare</td>
<td>79.46±1.11</td>
<td>82.74±1.90</td>
<td>82.27±1.86</td>
<td>82.18±1.45</td>
<td>82.55±1.75</td>
<td>83.40±1.67</td>
</tr>
<tr>
<td>9 german</td>
<td>74.70±1.33</td>
<td>72.30±1.57</td>
<td>73.10±1.54</td>
<td>73.90±1.85</td>
<td>72.20±1.23</td>
<td>73.70±2.02</td>
</tr>
<tr>
<td>10 glass</td>
<td>69.66±1.85</td>
<td>55.57±5.39</td>
<td>67.78±3.43</td>
<td>70.58±1.09</td>
<td>69.62±1.95</td>
<td>71.98±2.15</td>
</tr>
<tr>
<td>11 glass2</td>
<td>79.17±1.71</td>
<td>75.49±2.47</td>
<td>77.92±1.11</td>
<td>79.19±2.14</td>
<td>76.67±1.63</td>
<td>79.17±1.71</td>
</tr>
<tr>
<td>12 heart</td>
<td>81.48±3.26</td>
<td>82.22±2.46</td>
<td>82.33±2.48</td>
<td>82.22±2.96</td>
<td>81.11±3.77</td>
<td>81.85±2.83</td>
</tr>
<tr>
<td>13 hepatitis</td>
<td>91.25±1.53</td>
<td>91.25±4.68</td>
<td>91.25±2.50</td>
<td>92.50±1.25</td>
<td>86.25±4.15</td>
<td>90.00±4.24</td>
</tr>
<tr>
<td>14 iris</td>
<td>93.33±1.05</td>
<td>94.00±1.25</td>
<td>94.00±1.25</td>
<td>93.33±1.05</td>
<td>94.00±1.25</td>
<td>94.00±1.25</td>
</tr>
<tr>
<td>15 letter</td>
<td>74.96±0.61</td>
<td>75.02±0.61</td>
<td>85.86±0.49</td>
<td>88.02±0.46</td>
<td>77.70±0.59</td>
<td>75.36±0.61</td>
</tr>
<tr>
<td>16 lymphography</td>
<td>79.72±1.10</td>
<td>75.03±1.58</td>
<td>85.03±3.09</td>
<td>81.75±3.44</td>
<td>77.03±1.21</td>
<td>77.72±2.46</td>
</tr>
<tr>
<td>17 mofn-3-7-10</td>
<td>86.43±1.07</td>
<td>85.94±1.09</td>
<td>91.11±0.89</td>
<td>91.50±0.87</td>
<td>85.55±1.10</td>
<td>87.50±1.03</td>
</tr>
<tr>
<td>18 pima</td>
<td>75.51±1.63</td>
<td>75.00±1.22</td>
<td>75.52±1.27</td>
<td>75.39±1.51</td>
<td>75.13±1.52</td>
<td>74.86±2.61</td>
</tr>
<tr>
<td>19 satimage</td>
<td>81.75±0.86</td>
<td>59.20±1.10</td>
<td>87.20±0.75</td>
<td>87.00±0.75</td>
<td>83.15±0.84</td>
<td>82.05±0.86</td>
</tr>
<tr>
<td>20 segment</td>
<td>91.17±1.02</td>
<td>93.51±0.89</td>
<td>95.58±0.74</td>
<td>93.12±0.91</td>
<td>93.64±0.88</td>
<td>93.25±0.90</td>
</tr>
<tr>
<td>21 shuttle-small</td>
<td>98.34±0.29</td>
<td>99.17±0.21</td>
<td>99.53±0.15</td>
<td>99.02±0.22</td>
<td>99.17±0.21</td>
<td>99.28±0.19</td>
</tr>
<tr>
<td>22 soybean-large</td>
<td>91.29±0.98</td>
<td>58.54±4.84</td>
<td>92.17±1.02</td>
<td>91.46±0.99</td>
<td>92.00±1.11</td>
<td>92.89±1.01</td>
</tr>
<tr>
<td>23 vehicle</td>
<td>58.28±1.79</td>
<td>61.00±2.02</td>
<td>69.63±2.11</td>
<td>67.15±2.06</td>
<td>69.74±1.52</td>
<td>61.36±2.33</td>
</tr>
<tr>
<td>24 vote</td>
<td>90.34±0.86</td>
<td>94.94±0.46</td>
<td>93.56±0.28</td>
<td>94.71±1.00</td>
<td>95.63±0.43</td>
<td>94.71±0.59</td>
</tr>
<tr>
<td>25 waveform-21</td>
<td>77.89±0.61</td>
<td>69.45±0.67</td>
<td>78.38±0.60</td>
<td>78.36±0.60</td>
<td>74.70±0.63</td>
<td>76.53±0.62</td>
</tr>
</tbody>
</table>
Table 3. Experimental results describing the effect of smoothing parameters.

It is clear that, if two attributes are perfectly correlated, then the removal of one can only improve the performance of the naive Bayesian classifier. Problems arise, however, if two attributes are only partially correlated. In these cases the removal of an attribute may lead to the loss of useful information, and the selective naive Bayesian classifier may still retain both attributes. In addition, this wrapper-based approach is, in general, computationally expensive. Our experimental results (see Figure 6) show that the methods we examine here are usually more accurate than the selective naive Bayesian classifier as used by John and Kohavi (1997).

Work in the second category (Kononenko, 1991; Pazzani, 1995; Ezawa & Schuermann, 1995) are closer in spirit to our proposal, since they attempt to improve the predictive accuracy by removing some of the independence assumptions. The semi-naive Bayesian classifier (Kononenko, 1991) is a model of the form:

$$P(C, A_1, ..., A_n) = P(C) \cdot P(A_1 | C) \cdot ... \cdot P(A_k | C)$$  \hspace{1cm} (9)

where $A_1, ..., A_k$ are pairwise disjoint groups of attributes. Such a model assumes that $A_i$ is conditionally independent of $A_j$ if, and only if, they are in different groups. Thus, no assumption of independence is made about attributes that are in the same group. Kononenko's method uses statistical tests of independence to partition the attributes into groups. This procedure, however, tends to select large groups, which can lead to overfitting problems. The number of parameters needed to estimate $P(A_i | C)$ is $|\text{Val}(C)| \cdot (\prod_{A_j \in A_i} |\text{Val}(A_j)| - 1)$, which grows exponentially with the number of attributes in the group. Thus, the parameters...
Table 4. Experimental results of comparing tree-like networks with unrestricted augmented naive Bayes and
multinets.

<table>
<thead>
<tr>
<th>Data set</th>
<th>TAN*</th>
<th>ANB</th>
<th>CL*</th>
<th>MN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 australia</td>
<td>84.20+-1.24</td>
<td>86.81+-0.42</td>
<td>85.07+-1.31</td>
<td>86.52+-0.63</td>
</tr>
<tr>
<td>2 breast</td>
<td>96.92+-0.67</td>
<td>96.49+-1.09</td>
<td>97.07+-0.66</td>
<td>96.63+-0.95</td>
</tr>
<tr>
<td>3 chess</td>
<td>92.31+-0.82</td>
<td>94.18+-0.72</td>
<td>92.40+-0.81</td>
<td>96.34+-0.58</td>
</tr>
<tr>
<td>4 cleave</td>
<td>81.76+-0.33</td>
<td>80.08+-1.30</td>
<td>80.73+-1.40</td>
<td>81.76+-0.33</td>
</tr>
<tr>
<td>5 corral</td>
<td>96.06+-2.51</td>
<td>98.40+-1.60</td>
<td>99.23+-0.77</td>
<td>99.20+-0.80</td>
</tr>
<tr>
<td>6 crx</td>
<td>85.76+-1.16</td>
<td>86.37+-0.38</td>
<td>86.22+-1.14</td>
<td>86.37+-0.54</td>
</tr>
<tr>
<td>7 diabetes</td>
<td>75.52+-1.11</td>
<td>75.52+-1.06</td>
<td>74.74+-1.19</td>
<td>76.04+-0.75</td>
</tr>
<tr>
<td>8 flare</td>
<td>82.27+-1.86</td>
<td>82.84+-1.76</td>
<td>82.18+-1.45</td>
<td>82.65+-1.61</td>
</tr>
<tr>
<td>9 german</td>
<td>73.10+-1.54</td>
<td>73.20+-1.28</td>
<td>73.90+-1.85</td>
<td>72.20+-2.00</td>
</tr>
<tr>
<td>10 glass</td>
<td>67.78+-3.43</td>
<td>69.66+-1.85</td>
<td>70.58+-1.09</td>
<td>65.91+-1.63</td>
</tr>
<tr>
<td>11 glass2</td>
<td>77.92+-1.11</td>
<td>79.17+-1.71</td>
<td>79.19+-2.14</td>
<td>77.90+-1.21</td>
</tr>
<tr>
<td>12 heart</td>
<td>83.33+-2.48</td>
<td>82.59+-2.31</td>
<td>82.22+-2.96</td>
<td>83.70+-2.58</td>
</tr>
<tr>
<td>13 hepatitis</td>
<td>91.25+-2.50</td>
<td>88.75+-2.34</td>
<td>92.50+-1.25</td>
<td>90.00+-1.53</td>
</tr>
<tr>
<td>14 iris</td>
<td>94.00+-1.25</td>
<td>93.33+-1.05</td>
<td>93.33+-1.05</td>
<td>93.33+-1.05</td>
</tr>
<tr>
<td>15 letter</td>
<td>85.86+-0.49</td>
<td>76.60+-0.60</td>
<td>88.02+-0.46</td>
<td>80.10+-0.56</td>
</tr>
<tr>
<td>16 lymphography</td>
<td>85.03+-3.09</td>
<td>83.10+-2.19</td>
<td>81.75+-3.44</td>
<td>79.75+-0.97</td>
</tr>
<tr>
<td>17 mofn-3-7-10</td>
<td>91.11+-0.89</td>
<td>86.43+-1.07</td>
<td>91.50+-0.87</td>
<td>86.43+-1.07</td>
</tr>
<tr>
<td>18 pima</td>
<td>75.52+-1.27</td>
<td>74.74+-1.23</td>
<td>75.39+-1.51</td>
<td>76.30+-1.24</td>
</tr>
<tr>
<td>19 satimage</td>
<td>87.20+-0.75</td>
<td>80.50+-0.89</td>
<td>87.00+-0.75</td>
<td>77.10+-0.94</td>
</tr>
<tr>
<td>20 segment</td>
<td>95.58+-0.74</td>
<td>91.17+-1.02</td>
<td>93.12+-0.91</td>
<td>90.26+-1.07</td>
</tr>
<tr>
<td>21 shuttle-small</td>
<td>99.53+-0.15</td>
<td>98.91+-0.24</td>
<td>99.02+-0.22</td>
<td>98.97+-0.23</td>
</tr>
<tr>
<td>22 soybean-large</td>
<td>92.17+-1.02</td>
<td>92.18+-1.02</td>
<td>91.46+-0.99</td>
<td>87.01+-1.17</td>
</tr>
<tr>
<td>23 vehicle</td>
<td>69.63+-2.11</td>
<td>67.38+-1.38</td>
<td>67.15+-2.06</td>
<td>64.20+-2.57</td>
</tr>
<tr>
<td>24 vote</td>
<td>93.56+-0.28</td>
<td>89.66+-1.21</td>
<td>94.71+-1.00</td>
<td>90.11+-1.48</td>
</tr>
<tr>
<td>25 waveform-21</td>
<td>78.38+-0.60</td>
<td>77.75+-0.61</td>
<td>78.36+-0.60</td>
<td>76.85+-0.62</td>
</tr>
</tbody>
</table>
Example: data on death penalty

Data provided by the Georgia parole board.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>death</td>
<td>Did the defendant get the death penalty?</td>
</tr>
<tr>
<td>blkdef</td>
<td>Is the defendant black?</td>
</tr>
<tr>
<td>whtvict</td>
<td>Is the victim white?</td>
</tr>
<tr>
<td>aggcirc</td>
<td>Number of aggravating circumstances</td>
</tr>
<tr>
<td>stranger</td>
<td>Were victim and defendant strangers?</td>
</tr>
</tbody>
</table>

We have 100 observations.
Example in bnlearn package

```r
> death.small[1:5,]
  death blkdef whtvict aggcirc stranger
1 0 1 0 1 0
2 0 1 0 1 0
3 0 1 0 1 0
4 1 1 0 2 0
5 0 1 0 1 0

> summary(death.small)
  death blkdef whtvict aggcirc stranger
0:51 0:47 0:26 1:22 0:49
1:49 1:53 1:74 2:78 1:51

> death.nb <- naive.bayes("death",data=death.small)
> death.nb.pred <- predict(death.nb,death.small)
> table(death.small[,1],death.nb.pred)
  death.nb.pred
     0  1
  0 36 15
  1 12 37

> sum(diag(table(death.small[,1],death.nb.pred))/nrow(death.small))
[1] 0.73
```
Example in bnlearn package

```r
# fit TAN to death penalty data with "death" as class variable
> death.tan <- tree.bayes(death.small,"death")
# fit TAN parameters using maximum likelihood estimation ("mle")
> death.tan.fit <- bn.fit(death.tan,death.small,"mle")
# predict class on training sample
> death.tan.pred <- predict(death.tan.fit,death.small)
# make confusion matrix
> table(death.small[,1],death.tan.pred)
   death.tan.pred
      0 1
     0 32 19
     1  8 41
# compute accuracy
> sum(diag(table(death.small[,1],death.tan.pred))/nrow(death.small))
[1] 0.73
# plot the TAN structure
> plot(death.tan)
```
TAN for death penalty data
Example in bnlearn package

```R
# learn network structure with hill-climber
> death.hc <- hc(death.small)
> plot(death.hc)
> death.hc.fit <- bn.fit(death.hc,death.small,"mle")
> death.hc.pred <- predict(death.hc.fit,node="death",data=death.small)
> table(death.small[,1],death.hc.pred)
   death.hc.pred
       0 1
     0 20 31
     1  6 43
> sum(diag(table(death.small[,1],death.hc.pred))/nrow(death.small))
[1] 0.63
```
Network structure with hill climbing for death penalty data

- death
- blkdef
- stranger
- whtvict
- aggcirc

Ad Feelders (Universiteit Utrecht)
Data Mining
October 25, 2012
Death penalty and race of defendant are independent given race of victim.