On Robust Trimming of Bayesian Network Classifiers

YooJung Choi and Guy Van den Broeck
UCLA
Bayesian Network Classifiers

\[ \text{Pr}(C \mid \text{features}) \]
Bayesian Network Classifiers

\[ C_T(\text{features}) = \mathbb{I}(\Pr(C | \text{features}) \geq T) \]
Bayesian Network Classifiers

Can we make the same classifications with fewer features?

\[ C_T(\text{features}) = \mathbb{I}(\Pr(C \mid \text{features}) \geq T) \]
Why Classification Similarity?

To preserve classification behavior on individual examples

• Fairness
• Deployed classifiers
How to measure Similarity?

“Expected Classification Agreement”

$$ECA(\alpha, \beta) = \sum_{f} \mathbb{I}(C_T(f) = C_{T'}(f')) \cdot Pr(f)$$

What is the expected probability that a classifier $\alpha$ will agree with its trimming $\beta$?
Robust Trimming

\[
\max_{F' \subseteq F} \max_{T'} \text{ECA}(\alpha, (C, F', T')) \\
\text{s.t. } \text{cost}(F') \leq B
\]
Trimming Algorithm

Feature subset selection

$$\max_{F' \subseteq F} \max_{T'} ECA(\alpha, (C, F', T'))$$

“Maximum Achievable Agreement”

Search

Objective function
Trimming Algorithm

• Branch-and-Bound search
Trimming Algorithm

• Branch-and-Bound search
• Need a bound for MAA to prune subtrees
Upper-bound for MAA

“Maximum Potential Agreement”

\[ MPA_\alpha(F') = \sum_{f'} \max_c \sum_{f' \models f'} \mathbb{I}(C_T(f) = c) \Pr(f) \]

Maximum agreement between \( \alpha \) and a hypothetical function that maps \( f' \) to \( c \)
Maximum Potential Agreement

1. Upper-bounds the MAA
2. Monotonically increasing

Great for pruning!
Maximum Potential Agreement

1. Upper-bounds the MAA
2. Monotonically increasing
3. Generally easier to compute than MAA
4. Equal to MAA given some independence condition (e.g. Naïve Bayes)
Computing the MPA and MAA

Prior works based on knowledge compilation

\[ D \text{ Pr}(R_1 = + \mid D) \]
\[ + \quad 0.7 \]
\[ - \quad 0.2 \]

\[ \text{Pr}(D = +) \]
\[ 0.2 \]

\[ P_1 \Leftrightarrow D \land R_1 \]
\[ P_2 \Leftrightarrow D \land \neg R_1 \]
\[ P_3 \Leftrightarrow \neg D \land R_1 \]
\[ P_4 \Leftrightarrow \neg D \land \neg R_1 \]

\[ w(P_1) = 0.7 \]
\[ w(P_2) = 0.3 \]
\[ w(P_3) = 0.2 \]
\[ w(P_4) = 0.8 \]
\[ w(l) = 1.0 \text{ for all other literal } l \]

[Oztok, Choi, Darwiche 2016; C, Darwiche, VdB 2017]
Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Agreement</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>pima</td>
<td>Opt. ECA</td>
<td>0.9863</td>
</tr>
<tr>
<td></td>
<td>Opt. Acc.</td>
<td>0.9452</td>
</tr>
<tr>
<td>heart</td>
<td>Opt. ECA</td>
<td>0.9245</td>
</tr>
<tr>
<td></td>
<td>Opt. Acc.</td>
<td>0.9057</td>
</tr>
</tbody>
</table>
Evaluation

<table>
<thead>
<tr>
<th>F</th>
<th>FS-SDD</th>
<th>ECA-TRIM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time</td>
<td># Eval</td>
</tr>
<tr>
<td>bupa</td>
<td>0.044</td>
<td>21</td>
</tr>
<tr>
<td>pima</td>
<td>0.056</td>
<td>36</td>
</tr>
<tr>
<td>ident</td>
<td>0.128</td>
<td>129</td>
</tr>
<tr>
<td>anatomy</td>
<td>2.252</td>
<td>793</td>
</tr>
<tr>
<td>heart</td>
<td>7.346</td>
<td>1092</td>
</tr>
<tr>
<td>voting</td>
<td>819.163</td>
<td>6884</td>
</tr>
<tr>
<td>hepatitis</td>
<td>Timeout</td>
<td>43795</td>
</tr>
</tbody>
</table>

Branch-and-bound improves efficiency (even with extra upper-bound computations)
Evaluation

High information gain does not lead to high classification agreement

Information-theoretic measures unaware of changes in classification threshold
Thank you!

Questions?