Lazy Naive Credal Classifier

G. Corani    M. Zaffalon

IDSIA
Switzerland
giorgio{zaffalon}@idsia.ch

U’ 09 - Knowledge discovery from uncertain data
Naive Credal Classifier (NCC) (Corani and Zaffalon, JMLR 2008)

- NCC extends naive Bayes to imprecise probabilities.
- NCC returns *indeterminate* classifications (more than one class) on *hard* instances, i.e., for which the evidence does not smooth the effect of the chosen prior (e.g., small samples);
- Naive Bayes is unreliable on such hard instances.
Two issues of NCC

- The naive assumption (statistical independence of the features given the class) can be unrealistic in some domains.
- A sometimes excessive indeterminacy.
- Lazy Naive Credal Classifier (LNCC) is designed to overcome both issues.
Lazy Classifiers

Lazy classifiers do not learn until there is an instance to classify (query).

Then:

1. they rank the instances of the training set according to the distance from the query;
2. a local classifier is trained on the $k$ closest instances ($k$ is named bandwidth) and issues the classification;
3. the local classifier is discarded, while the training set is kept in memory to answer future queries.

Open question: how to select the bandwidth?
The bandwidth is increased until LNCC is determinate, i.e., until the evidence smooths the effect of the choice of the prior.

**Pseudo-code**

- $k=25$;
- \texttt{lncc.train}(k);
- while (lncc indeterminate OR $k == \text{trainingSet.size}$)
  - $k=k+20$;
  - \texttt{lncc.train}(k);
- end
Why a lazy NCC (LNCC)?

- working locally *reduces the bias* due to the naive assumption (J. Friedman, 1997; Frank et al., UAI 2003);
- to *increase determinacy*, thanks the design of the *bandwidth selector*.
Comparing credal classifiers: d-acc

A classifier is *accurate* on a certain instance if its output includes the correct class.

- Discounted accuracy (borrowed from multi-label classification):

\[
d-\text{acc} = \frac{1}{N} \sum_{i=1}^{N} \frac{(\text{accurate})_i}{|Z_i|}
\]

where \(|Z_i|\) is the number of classes returned on the \(i\)-th instance.

- \(d\)-acc entails some arbitrariness: why not discounting on \(|Z_i|^2\)?
- The \(d\)-acc of two credal classifiers can be compared via \(t\)-test.
Comparing credal classifier: a new rank test

On each instance, we rank the two credal classifiers $CR_1$ and $CR_2$:

<table>
<thead>
<tr>
<th></th>
<th>CR1</th>
<th>CR2</th>
<th>winner</th>
</tr>
</thead>
<tbody>
<tr>
<td>accurate</td>
<td>not accurate</td>
<td></td>
<td>$CR_1$</td>
</tr>
<tr>
<td>accurate</td>
<td>accurate</td>
<td>$</td>
<td>Z_i</td>
</tr>
<tr>
<td>accurate</td>
<td>accurate</td>
<td>$</td>
<td>Z_i</td>
</tr>
<tr>
<td>inaccurate</td>
<td>inaccurate</td>
<td></td>
<td>tie</td>
</tr>
</tbody>
</table>

- Wins, ties and losses are transformed into ranks and are then analyzed via a non-parametric test.
- The rank test avoids the arbitrariness of d-acc but, using less pieces of information, can be less sensitive.
Experiments

- Comparison of LNCC and NCC on 36 data sets.

<table>
<thead>
<tr>
<th></th>
<th>LNCC wins</th>
<th>ties</th>
<th>NCC wins</th>
</tr>
</thead>
<tbody>
<tr>
<td>d-acc</td>
<td>19</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>rank test</td>
<td>15</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>cross-check</td>
<td>15</td>
<td>20</td>
<td>1</td>
</tr>
</tbody>
</table>
Why does LNCC outperform NCC?

- On large data sets the improvement is due to the reduced bias.

<table>
<thead>
<tr>
<th>Data set</th>
<th>instances</th>
<th>d-acc (NCC)</th>
<th>ΔLNCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>letter</td>
<td>20000</td>
<td>86.5</td>
<td>+12.1</td>
</tr>
<tr>
<td>nursery</td>
<td>12960</td>
<td>95.8</td>
<td>+5.6</td>
</tr>
<tr>
<td>optdigits</td>
<td>5620</td>
<td>93.9</td>
<td>+1.9</td>
</tr>
<tr>
<td>pendigits</td>
<td>10992</td>
<td>94.3</td>
<td>+6.3</td>
</tr>
<tr>
<td>waveform</td>
<td>5000</td>
<td>84.0</td>
<td>+4.1</td>
</tr>
</tbody>
</table>

- In other cases there is a considerable improvement of determinacy.
Conclusions

- Thanks to locality and to the design of the bandwidth selector, LNCC overcomes two problems of NCC:
  - the bias of the naive assumption;
  - the (sometimes) excessive indeterminacy;
- We have proposed two metrics for comparing credal classifiers;
- Experiments on 36 data sets show a clear improvement of LNCC over NCC.