Naive Credal Classifier 2: an extension of Naive Bayes for delivering robust classifications

G. Corani  M. Zaffalon

IDSIA
Switzerland
giorgio{zaffalon}@idsia.ch
DMIN ’08

Introducing NCC2

1. Background
2. Credal classifiers
3. NCC2

Experimental Results

1. Setup and indicators
2. Indeterminate classifications vs posterior probabilities

Conclusions

Naive Bayes Classifier (NBC)

- Naive assumption (statistical indep. of the features given the class):

\[ \theta_{y_1, y_2, \ldots, y_k} \propto \theta_0 \prod_{i=1}^{k} \theta_{y_i|c} \]

Probability computation

\[ \theta_{\text{POST}} \propto \theta_{\text{Likelihood}} \theta_{\text{Prior}} \]

- Maximum likelihood estimators are for instance \( \hat{\theta}_c = n(c)/N \) and \( \hat{\theta}_{y_i|c} = n(y_i|c)/n(c) \).
- The choice of any specific prior introduces necessarily some subjectivity.

NBC and prior sensitivity

- NBC computes a single posterior distribution.
- However, the most probable class might depend on the chosen prior, especially on small data sets.
- Prior-dependent classifications might be fragile.
- Solution via set of probabilities:
  - Robust Bayes Classifier (Ramoni and Sebastiani, 2001)
  - Naive Credal Classifier (Zaffalon, 2001)
Introducing NCC2

Naive Credal Classifier (NCC) (Zaffalon, 2001)

- Extends Naive Bayes to imprecise probabilities; it specifies a set of priors by adopting the Imprecise Dirichlet Model.
- The set of priors is turned into a set of posteriors via Bayes’ rule.
- NCC returns the classes that are non-dominated within the set of posteriors.

Test of dominance and indeterminate classifications

Definition

Class \( c_1 \) dominates \( c_2 \) if \( P(c_1) > P(c_2) \) in any distribution of the posterior credal set.

If no class dominates \( c_1 \), then \( c_1 \) is non-dominated.

- If there are more non-dominated classes, NCC returns all of them: the classification is indeterminate.
- NCC becomes indeterminate on the instances whose classification would be prior-dependent with NBC.
- Indeterminate classifications proofed to be viable in real world case studies (e.g., dementia diagnosis).

Incomplete data sets

- Most classifiers (including NBC) ignore missing data.
- This is correct only if data are missing-at-random (MAR).
- It is not possible to test the MAR hypothesis on the incomplete data.
- However, ignoring Non-MAR missing data can lead to unreliable conclusions.
- Missing data can be MAR for some features but not for some others; or can be MAR only in training and not in testing (or vice versa).

NCC2: NCC with conservative treatment of missing data

NCC2 receives the declaration of which features have Non-MAR missing data and at which stage (learning, testing or both).

- NCC2 ignores MAR missing data.
- NCC2 deals conservatively with Non-MAR missing data.

Conservative treatment of missing data (learning set)

- All possible completions of missing data are seen as possible.
- A set of likelihoods is computed.
- A set of posteriors is computed from a set of priors and a set of likelihoods.
- The conservative treatment of missing data can generate additional indeterminacy.
NCC2: NCC with conservative treatment of missing data (II)

**Conservative treatment of missing data in the instance to be classified**
- All possible completions of missing data are seen as possible, thus giving rise to several virtual instances.
- Test of dominance: $c_1$ should dominate $c_2$ on all the virtual instances.
- A procedure allows to find out the dominance relationships without actually building the virtual instances.
- Conservative treatment of missing data in the instance to classify can generate additional indeterminacy.

By adopting imprecise probabilities, NCC2 is designed to be robust to:
- prior specification, especially critical on small data sets;
- Non-MAR missing data, critical on incomplete data sets.
- However, excessive indeterminacy is undesirable.

**What to expect from indeterminate classifications**
- To preserve NCC2 reliability, avoiding too strong conclusions (a single class) on doubtful instances.
- To convey sensible information, dropping unlikely classes.

**Indicators of performance for NCC2**
- **determinacy** (% of determinate classifications);
- **single-accuracy** (% of determ. classification that are accurate);
- **set-accuracy** (% of indeterm. classifications that contain the true class);
- **size of indeterminate output**, i.e., avg. number of classes returned when indeterminate.

**Indicators for comparing NBC and NCC2**
- NBC(NCC2 D): accuracy of NBC on instances determinately classified by NCC2.
- NBC(NCC2 I): accuracy of NBC on instances indeterminately classified by NCC2.
- We expect NBC(NCC2 D) > NBC(NCC2 I).
Introducing NCC2

Experimental setting

- 18 UCI complete data sets (numerical features are discretized).

MAR setting

- Each observation (apart from the class) is turned into missing with 5% probability.
- All features declared MAR for NCC2.

Non-MAR setting

- Split the categories of each feature into two halves.
- Turn into missing with probability 5% the observations falling only in the first half values.
- All features declared Non-MAR for NCC2.

Experimental Results

Results on 18 UCI data sets (10 runs of 10 folds c-v)

- MAR setup: 5% missing data generated via a MAR mechanism; all features declared as MAR to NCC2.
- Non-MAR setup: 5% missing data generated via a Non-MAR mechanism; all features declared as Non-MAR to NCC2.
- Average NBC accuracy under both settings: 82%.

**NBC vs NCC2**

- NBC (NCC2 D): 85% (95%)
- NBC (NCC2 I): 36% (69%)

On each data set and setup: NBC (NCC2 D) > NBC (NCC2 I)

**NCC2**

- determinacy: 95% (52%)
- single accuracy: 85% (95%)
- set-accuracy: 85% (96%)
- imprecise output size: \( \approx 33\% \) of the classes

Indeterminate classifications do preserve the reliability of NCC2!

**NBC Probabilities vs indeterminate classifications**

(MAR setup, average over all data sets)

- Higher posterior probability of NBC \( \rightarrow \) higher NCC2 determinacy.
- At any level of posterior probability, NBC(NCC2 D) > NBC(NCC2 I).
- Striking drop on the instances classified confidently by NBC.

Non-MAR missing data lead to indeterminate classifications even if the probability computed by NBC is high.

At any level of posterior probability, NBC(NCC2 D) > NBC(NCC2 I).
Summary

- NCC2 extends Naive Bayes to imprecise probabilities, to robustly deal with small data sets and missing data.
- NCC2 becomes indeterminate on instances whose classification is doubtful indeed.
- Indeterminate classifications preserve the classifier' reliability while conveying sensible information.
- Bibliography, software and manuals: see www.idsia.ch/~giorgio/jncc2.html
- Software with GUI to arrive soon!