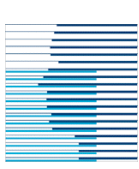


A TAN classifier based on the Extreme Dirichlet Model

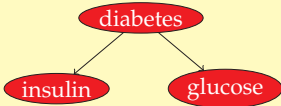


G. Corani, C. De Campos and S. Yi, *IDSIA, Switzerland*

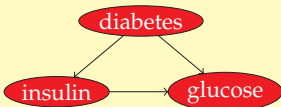
giorgio{cassio,yi}@idsia.ch

Tree-augmented networks (TAN)

- Within naive Bayes, the only parent of a feature is the class node.
- Within TAN (Friedman et al., 1997), the parents of a feature are the class and *at most another feature*, thus also modelling correlated features.



(a) Naive Bayes



(b) TAN

Credal classifiers

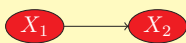
- Credal classifiers model prior *near-ignorance* by the Imprecise Dirichlet Model (IDM).
- They return the *non-dominated* classes in posterior credal set K .
- Class c' dominates c'' iff:

$$\min_{p \in K} (p(c'|\mathbf{y}) - p(c''|\mathbf{y})) > 0$$
 where \mathbf{y} denotes the value of the features.
- There can be more non-dominated classes (*indeterminate classification*).
- Usually, the accuracy of Bayesian classifiers drops on the instances indeterminately classified by their credal counterparts.

Variants of the IDM

- We consider three variants of the IDM: the *global*, the *local* and the *extreme* (EDM; Cano et al., 2007).

EXAMPLE



- The interval of X_1 is estimated identically by all IDMs as $\left[\frac{n_{x_1}}{n+s}, \frac{n_{x_1}+s}{n+s} \right]$.
- The probability interval of X_2 is estimated as follows:

$$\text{Global IDM} \quad \left[\frac{n_{x_1 x_2}}{n_{x_1} + \alpha_{x_1}}, \frac{n_{x_1 x_2} + \alpha_{x_1}}{n_{x_1} + \alpha_{x_1}} \right]$$

$$\text{Local IDM} \quad \left[\frac{n_{x_1 x_2}}{n_{x_1} + s}, \frac{n_{x_1 x_2} + s}{n_{x_1} + s} \right]$$

$$\text{EDM} \quad \left[\frac{n_{x_1 x_2}}{n_{x_1} + \alpha_{x_1}} \right]; \left[\frac{n_{x_1 x_2} + \alpha_{x_1}}{n_{x_1} + \alpha_{x_1}} \right]$$

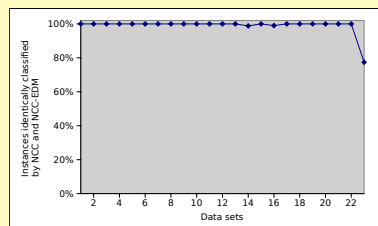
- $\alpha_{x_1 x_2}$ is constrained by $\sum_{x_2} \alpha_{x_1 x_2} = \alpha_{x_1}$ in the global IDM, but freely ranges between 0 and s in the local IDM.

IDM for classification

- The global IDM has been computed only for the naive credal classifier (NCC) (Corani and Zaffalon, 2007).
- The local IDM is easier to compute, but generates unnecessary indeterminacy.
- The EDM restricts the global IDM to its extreme distributions, thus simplifying the computation.
- The EDM treats the s hidden instances as s rows of missing data, whose actual values are unknown yet identical across the s instances.

The EDM has still to be tested in classification problems.

- We compared the NCC with EDM (NCC-EDM) and the NCC with global IDM.
- NCC searches the minimum of $p(c'|\mathbf{y}) - p(c''|\mathbf{y})$ in $(0, s)$ while NCC-EDM only evaluates this difference in 0 and s .



- Only on audiology NCC and NCC-EDM return 22% different classification; in fact, most features in this data set have *very* skewed distributions (e.g., 225,1).
- EDM appears however as a reliable approximation of the global IDM.

TANC

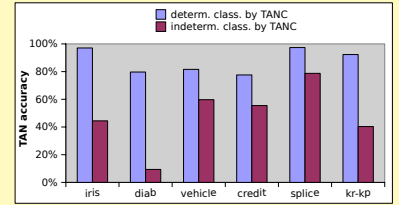
- A credal TAN was proposed by Zaffalon et al., (2003); it is referred to as TANC*.
- TANC* is based on the *local* IDM to keep feasible the computation and assumes missing data to be MAR.
- TANC* is reliable but returns many indeterminate classifications;

We propose TANC that:

- is based on the EDM, to avoid the unnecessary indeterminacy of the local IDM;
- treats missing data as nonMAR, by considering as possible all the replacements for missing data, thus computing a set of likelihoods.
- although the possible likelihoods increase exponentially with the number of missing values, the computational complexity of TANC does not.

Experiments

- Several UCI data sets; 10-folds cross-validation.
- The accuracy of the standard TAN sharply drops on the instances indeterminately classified by TANC.



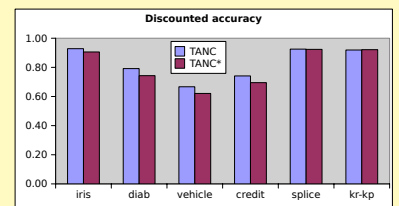
TANC vs. TAN

- We compare TANC against TANC*, by means of *discounted-accuracy*:

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

where $|Z_i|$ is the number of classes returned on the i -th instance and accurate_i denotes whether the output of the classifier includes the real class.

- TANC achieves higher d-acc than TANC*, mainly due to the removal of unnecessary indeterminacy.



TANC vs TANC*

Missing data

- The determinacy of TANC quickly decreases with the number of missing data; the TAN structure leads to higher indeterminacy than the naive one.

References

- G. Corani, M. Zaffalon, *Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2*, J. Mach. Learning Research, (9), 581–621, 2008.
- A. Cano, M. Gómez-Olmedo, S. Moral, *Credal nets with probabilities estimated with an extreme imprecise Dirichlet model*, ISIPTA 07, 57–66
- M. Zaffalon, E. Fagiouli, *Tree-based credal networks for classification*, Reliable Computing, 9(6), 487–509, 2003
- N. Friedman, D. Geiger, M. Goldszmidt, *Bayesian Networks Classifiers*, Machine Learning, 29(2), 131–163, 1997