

A TAN classifier based on the Extreme Dirichlet Model

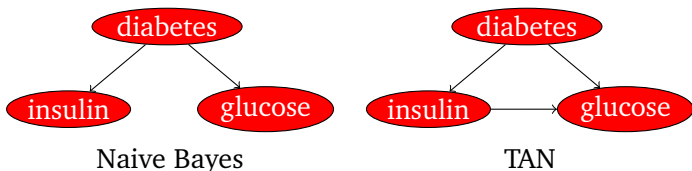
G. Corani C. De Campos S. Yi

IDSIA
Switzerland
giorgio{cassio, yi}@idsia.ch

ISIPTA 09

Tree-augmented networks (TAN)

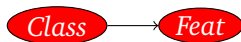
- Within naive Bayes, the only parent of a feature is the class node.
- Within TAN (Friedman et al., 1997), the parents of a feature are the class and *at most another feature*, thus also modelling correlated features.



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 .
- Denoting as \mathbf{y} the value of the features, class c' dominates c'' iff:
$$\min_{p \in K} [p(c', \mathbf{y}) - p(c'', \mathbf{y})] > 0$$
- If there are more non-dominated classes, the classification is *indeterminate*.

IDM in classification



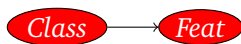
Local IDM:

- The credal sets K_{Class} and $K_{Feat|Class}$ are separately specified.
- The joint credal set is the strong extension of the local credal sets.
- Can produce wide intervals.

Classifiers:

- The very first NCC (naive credal classifier): Zaffalon, ISIPTA 99.
- TANC*, a credal TAN (Zaffalon et al., 2003).
- Both classifiers are reliable, but often indeterminate.

IDM in classification (2)



Global IDM:

- The joint credal set $K_{Class,Feat}$ is specified.
- The credal sets K_{Class} and $K_{Feat|Class}$ are linked by constraints.
- This makes difficult the computation of upper and lower probabilities.

Classifiers:

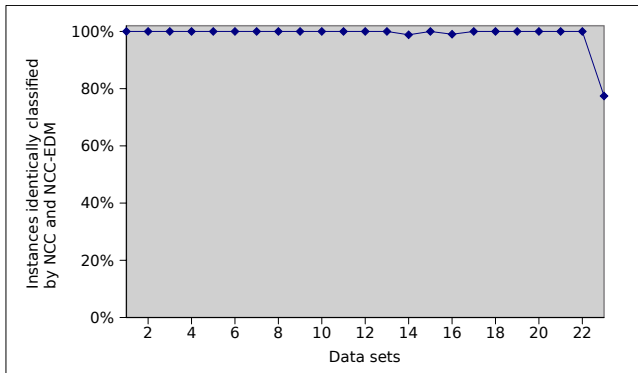
- The NCC of ISIPTA 2001, that obtained greater determinacy than the previous version.
- A credal TAN based on the global IDM could not be exactly computed.

The extreme IDM (Cano et al., ISIPTA 07)

- Restricts the global IDM to its extreme distributions.
- Practical interpretation:
 - the class and the feature have fixed values across the s hidden instances;
 - the credal set models that we are ignorant about such fixed values.
- Applied to credal networks, it can produce smaller intervals than the global IDM.
- Yet, *"it is necessary to test the behaviour of the EDM in real classification problems and to study the differences with the global IDM."*

NCC: global IDM vs EDM

- We have compared NCC with global IDM and with EDM.
- How often do the two models return the same classification?



- NCC-EDM is a close approximation of NCC.

The novel credal TAN (TANC)

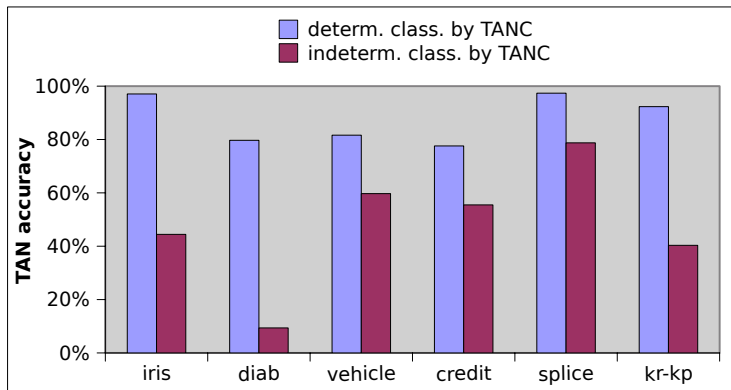
- TANC* (Zaffalon et al., (2003)):
 - based on the *local* IDM to make the computation feasible;
 - assumes missing data to be *missing at random* (MAR);
 - reliable but returns many indeterminate classifications.

The novel TANC:

- *based on the EDM;*
- *treats missing data without assuming MAR, by computing a set of likelihoods;*
- *while the likelihoods increase exponentially with the missing values, the computational complexity of TANC does not.*

TANC vs. Bayesian TAN

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



TANC vs TANC*: how to compare credal classifiers?

- **Note:** A classifier is *accurate* on a certain instance if its output includes the correct class.
- Discounted accuracy (borrowed from multi-label classification):

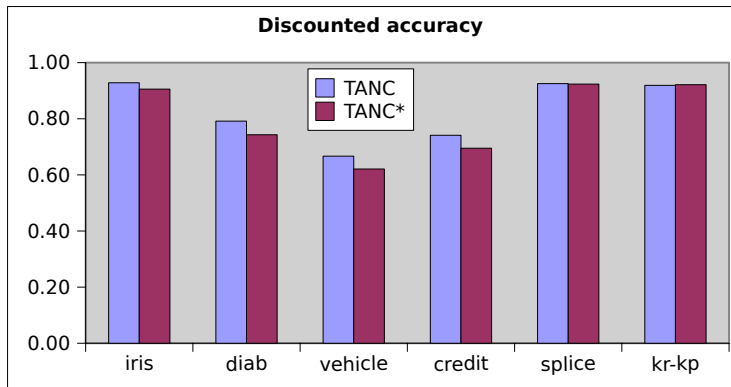
$$\text{d-acc} = \frac{1}{N} \sum_{i=1}^N \frac{(\text{accurate})_i}{|Z_i|}$$

where $|Z_i|$ is the number of classes returned on the i -th instance.

- d-acc entails some arbitrariness: why not discounting on $|Z_i|^2$?

TANC vs. TANC*

- TANC achieves higher d-acc than TANC*, mainly due to the removal of unnecessary indeterminacy.
- Splice and kr-kp contain around 3200 instances and the model of prior ignorance does not make a difference.



Conclusions

- EDM is a viable model of prior ignorance for classification.
- TANC improves TANC* by removing some unnecessary indeterminacy.
- Preliminary results with missing data: the indeterminacy quickly increases with the number of missing values.
- NCC is open source: add your own credal classifier to the package!
(www.idsia.ch/~giorgio/jncc2.html)