OBTAINING BEST PARAMETER VALUES FOR ACCURATE CLASSIFICATION

Frans Coenen and Paul Leng

Department of Computer Science,
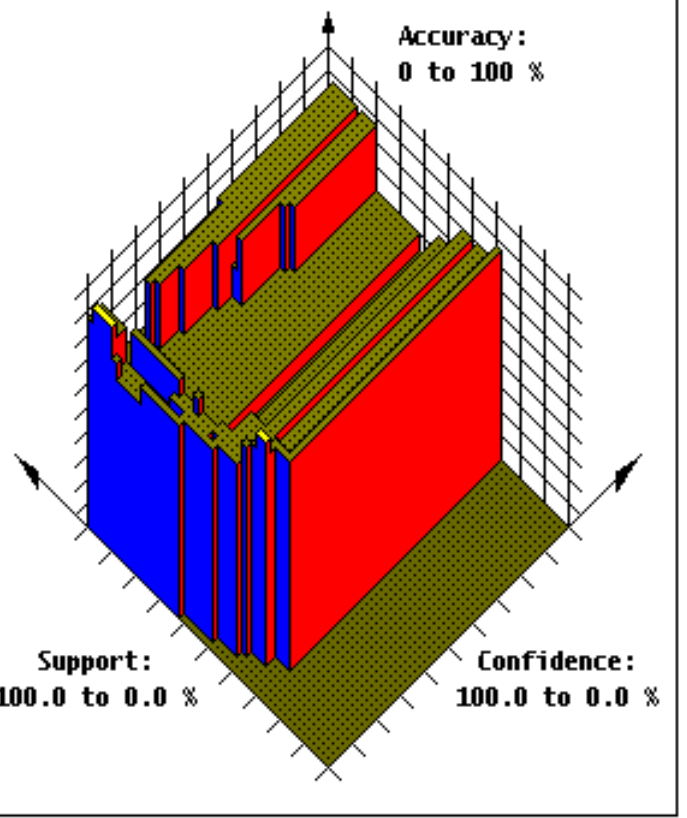
The University of Liverpool,
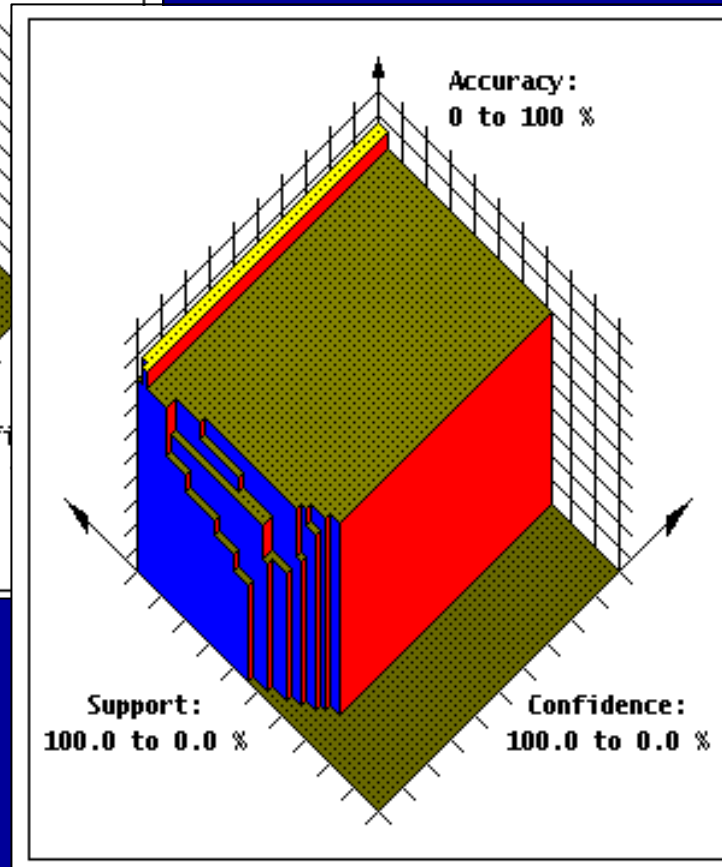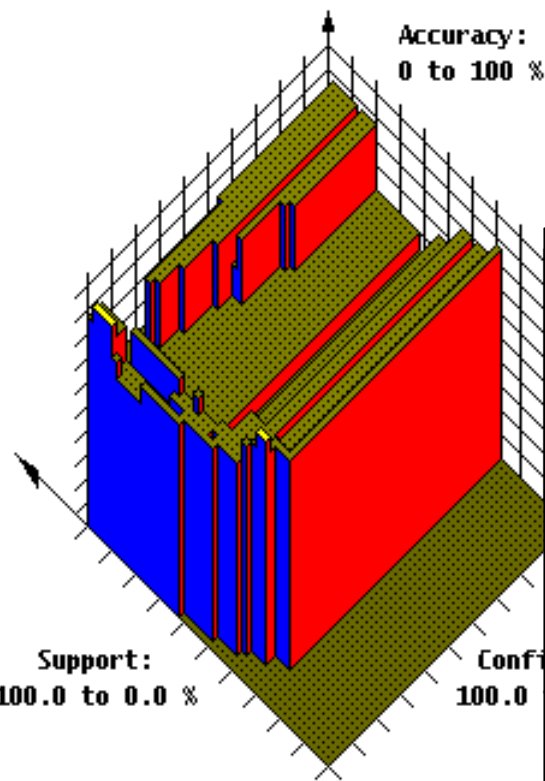
Liverpool, UK

{frans,phl}@csc.liv.ac.uk

PROPOSAL

- Many Classification Association Rule Mining (CARM) algorithms use various threshold values to generate their classifier.

- Q1: Are there "best" threshold value settings?

- Q2: How do we find these "best" values in a computationally effective/efficient manner?

Accuracy:
0 to 100 %

CBA (Liu et al., 1998)

Support:
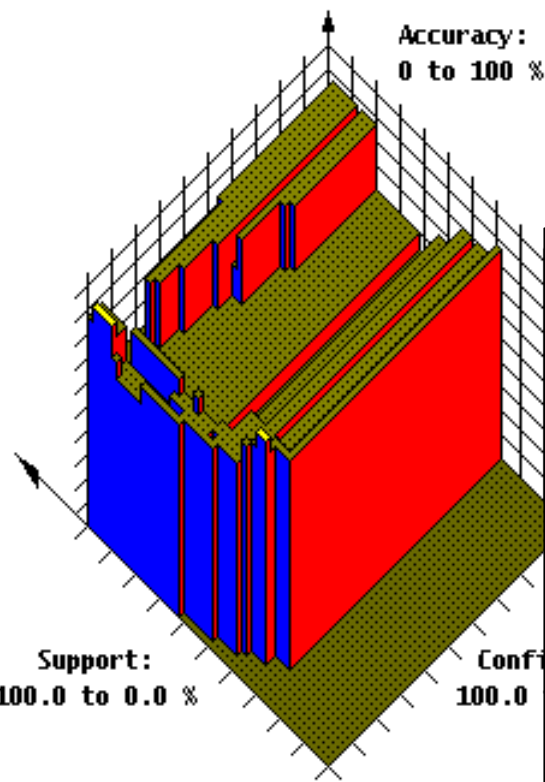100.0 to 0.0 %

Confidence:
100.0 to 0.0 %

Hepatitis

CBA (Liu et al., 1998)

CMAR (Li et al. 2001)

Hepatitis
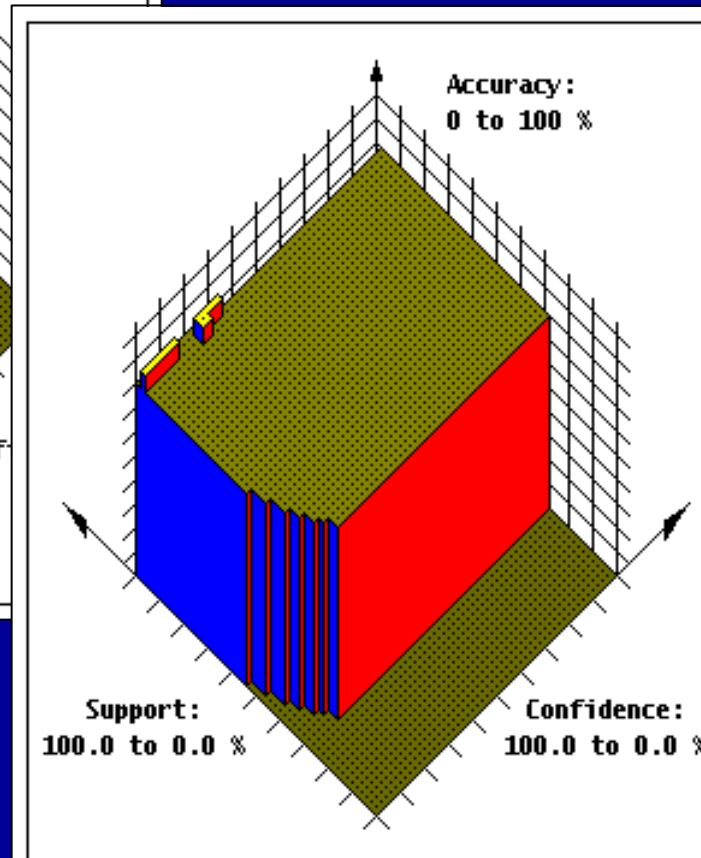
CBA (Liu et al., 1998)

TFPC (Coenen and Leng, 2005)
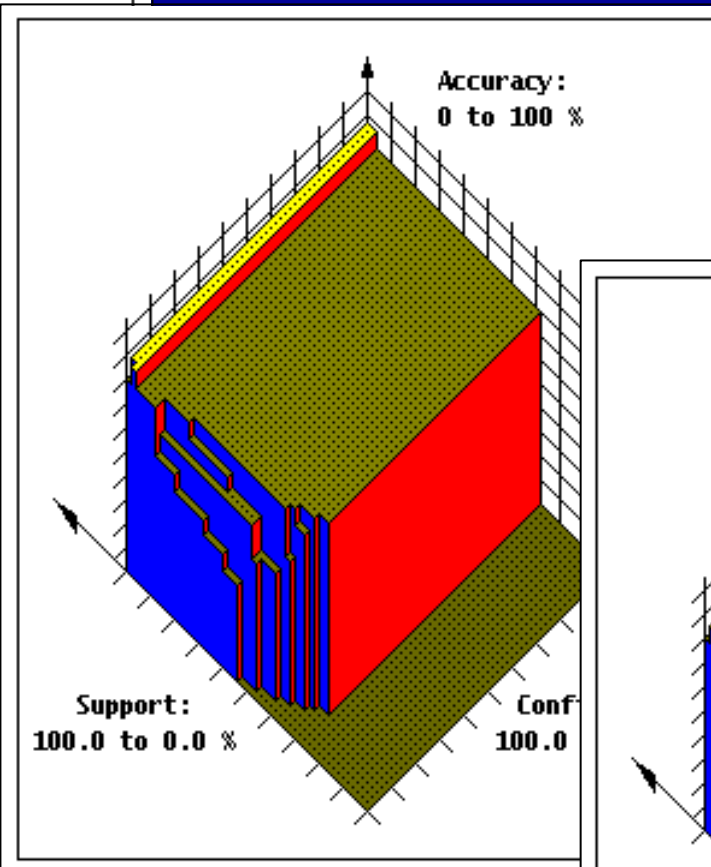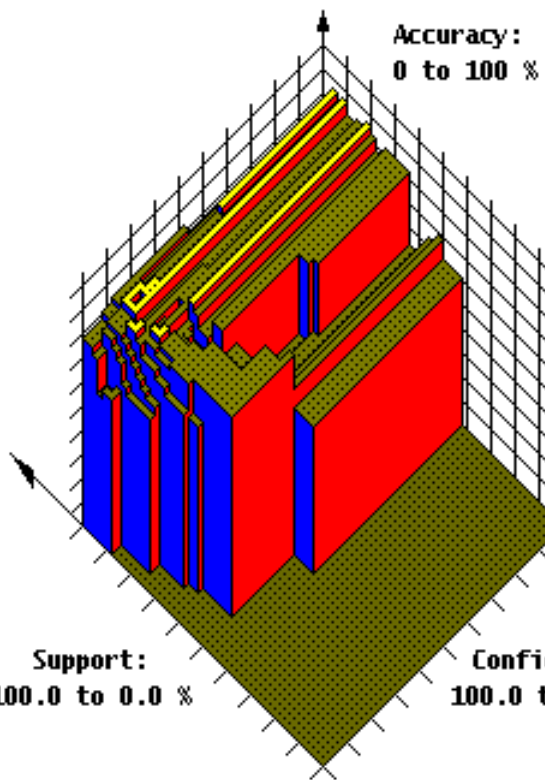
CMAR (Li et al. 2001)

Hepatitis

CBA (Liu et al., 1998)

TFPC
(Coenen and
Leng, 2005)
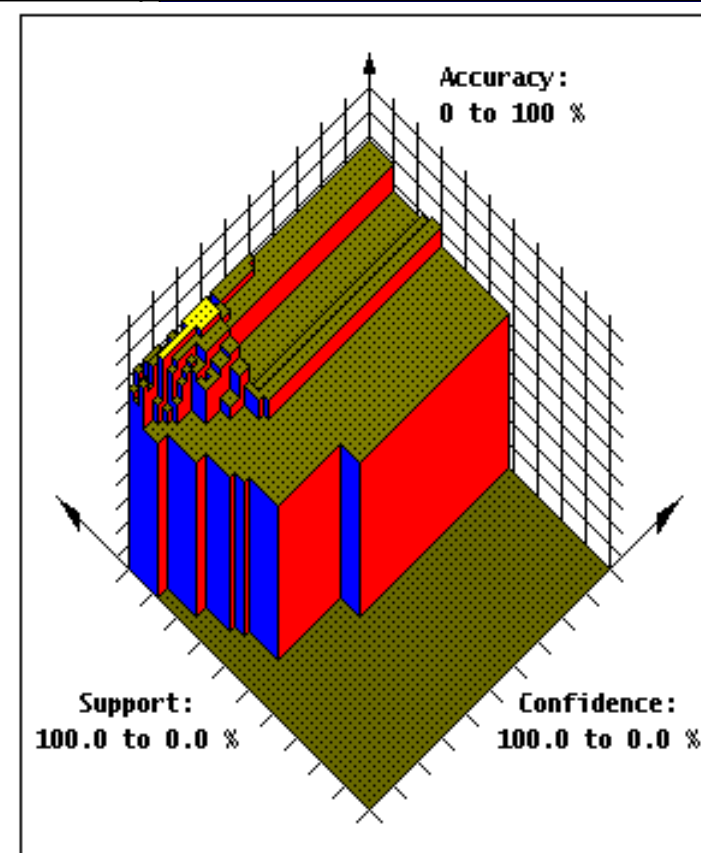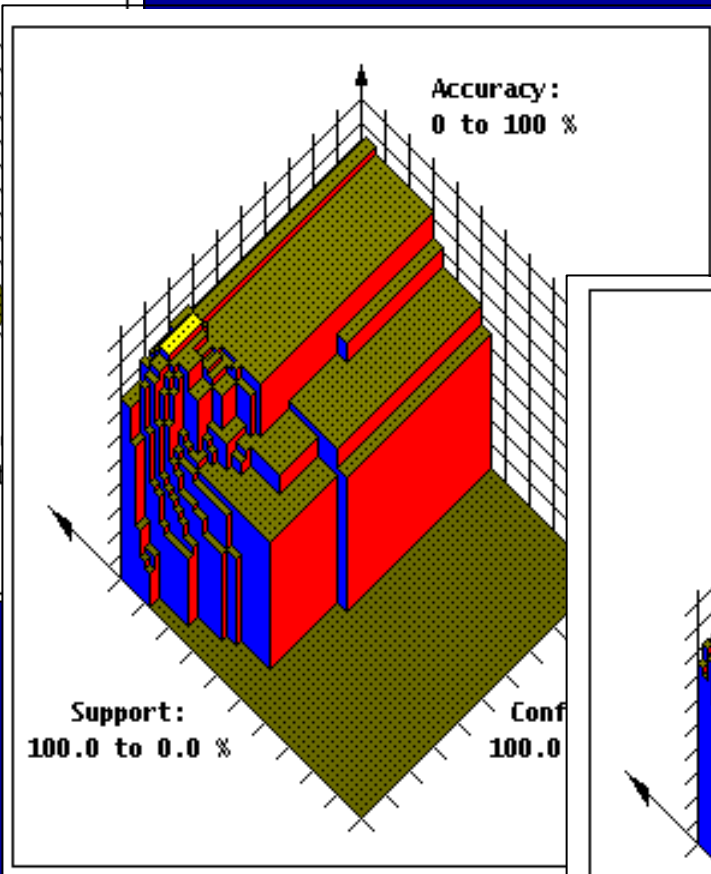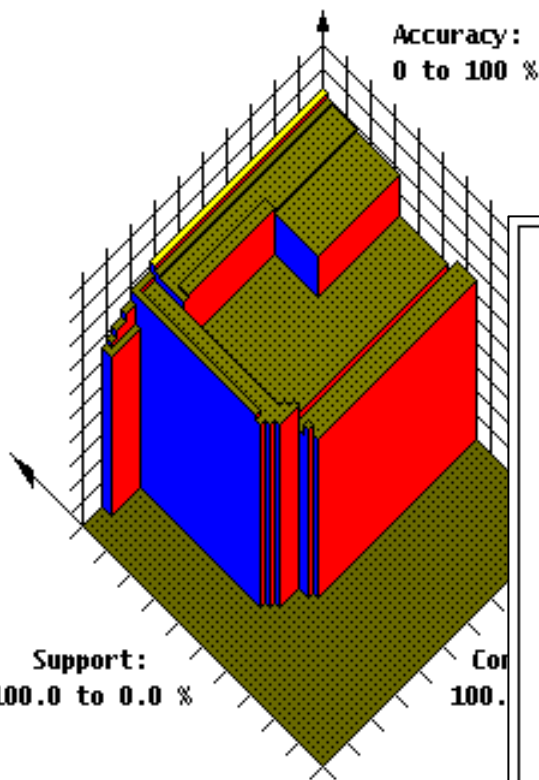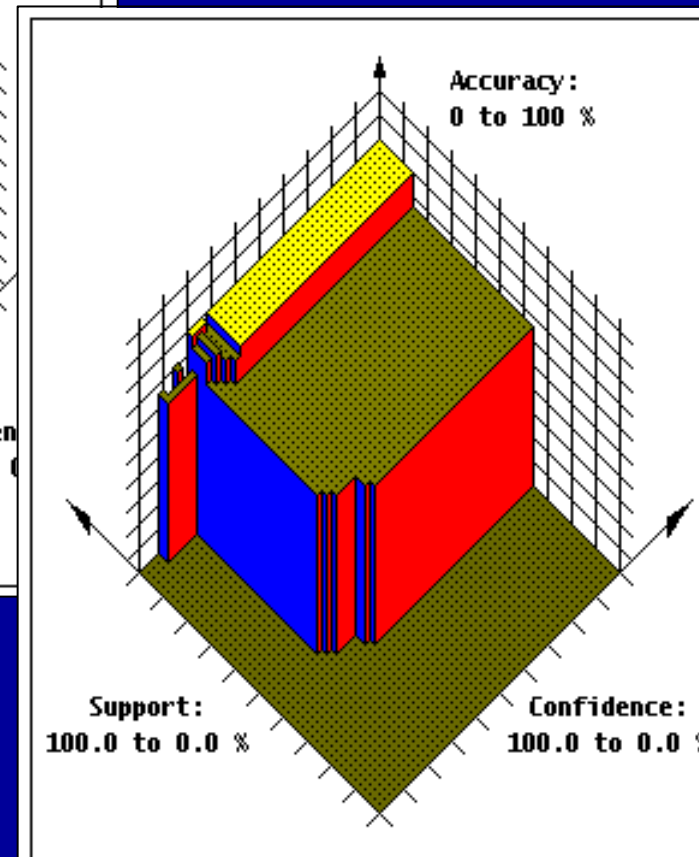
CMAR (Li
et al. 2001)
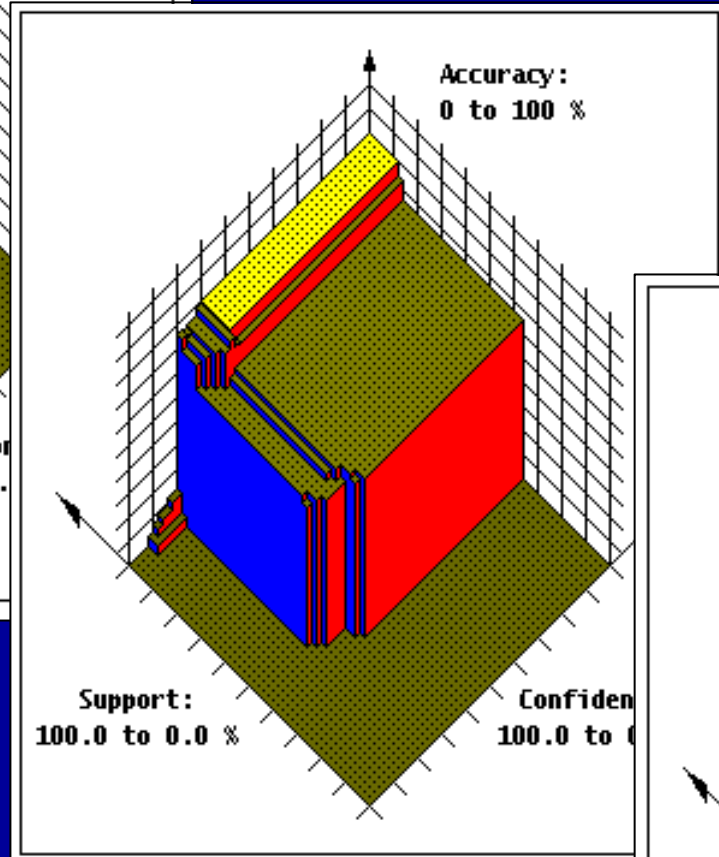
Horse
Collic

CBA (Liu et al., 1998)

TFPC (Coenen and Leng, 2005)

Horse Collic

CMAR (Li et al. 2001)

HILL CLIMBING

- Hill climbing techniques, coupled with heuristics to avoid local maxima, may be usefully employed to achieve the desired result.

- However, there are computational efficiency considerations!

- The TFPC (Total From Partial Classification) algorithm achieves this.

EXPERIMENTS

- Used CBA, CMAR and TFPC with 25 datasets from the UCI repository.

- Performing hill-climbing with TFPC is in many cases faster than coverage analysis with CMAR or CBA.

- In 13 of the 25 case TPFC was the fastest procedure to obtain classification rules (it was only markedly worse in cases such as `chess` and `letRecog`, where the other methods failed to identify the rules necessary for good classification accuracy.

CONCLUSIONS

Results suggest that:

- TFPC is a fast CARM algorithm.

- TFPC with hill-climbing is an effective way of generating a "best" classifier which is often less costly than other methods.