Threshold Tuning for Improved Classification Association Rule Mining

Frans Coenen[1], Paul Leng[1] and Lu Zhang[2]

1. Department of Computer Science, The University of Liverpool, Liverpool, UK
2. Department of Computer Science and Technology, Peking University, Beijing, P.R. China

PRESENTATION OVERVIEW

- Generation of Classification Association Rules (CARs): $X \Rightarrow c$
- Wish to avoid the "over-fit and prune" cycle so as to enhance computational efficiency.
- Propose an algorithm, TFPC, that generates CARs without the need to prune.
- TFPC makes use of the "support and confidence" framework and is thus sensitive to the selected thresholds.
- A threshold tuning mechanism is thus also presented.

THE "OVER-FIT AND PRUNE STRATEGY"

Typified by algorithms such as:

- CBA (Liu et al 1998)
- CMAR (Li et al 2001)

(Apriori/FPgrowth algorithm used to generate CARs, coverage analysis to prune)

TFPC (Total From Partial Classification)

- The intuition behind TFPC is that if we find a classification rule, $X \Rightarrow c$, that meets the user supplied confidence threshold, there is no need to continue processing to find further rules (that have higher confidence) with:
  - the consequent c and
  - antecedents which are supersets of X.
- Note also the TFPC operates in an Apriori manner therefore rules with small antecedents are generated before rules with large antecedents.
- In this manner TFPC avoids the “over-fit and prune” cycle.

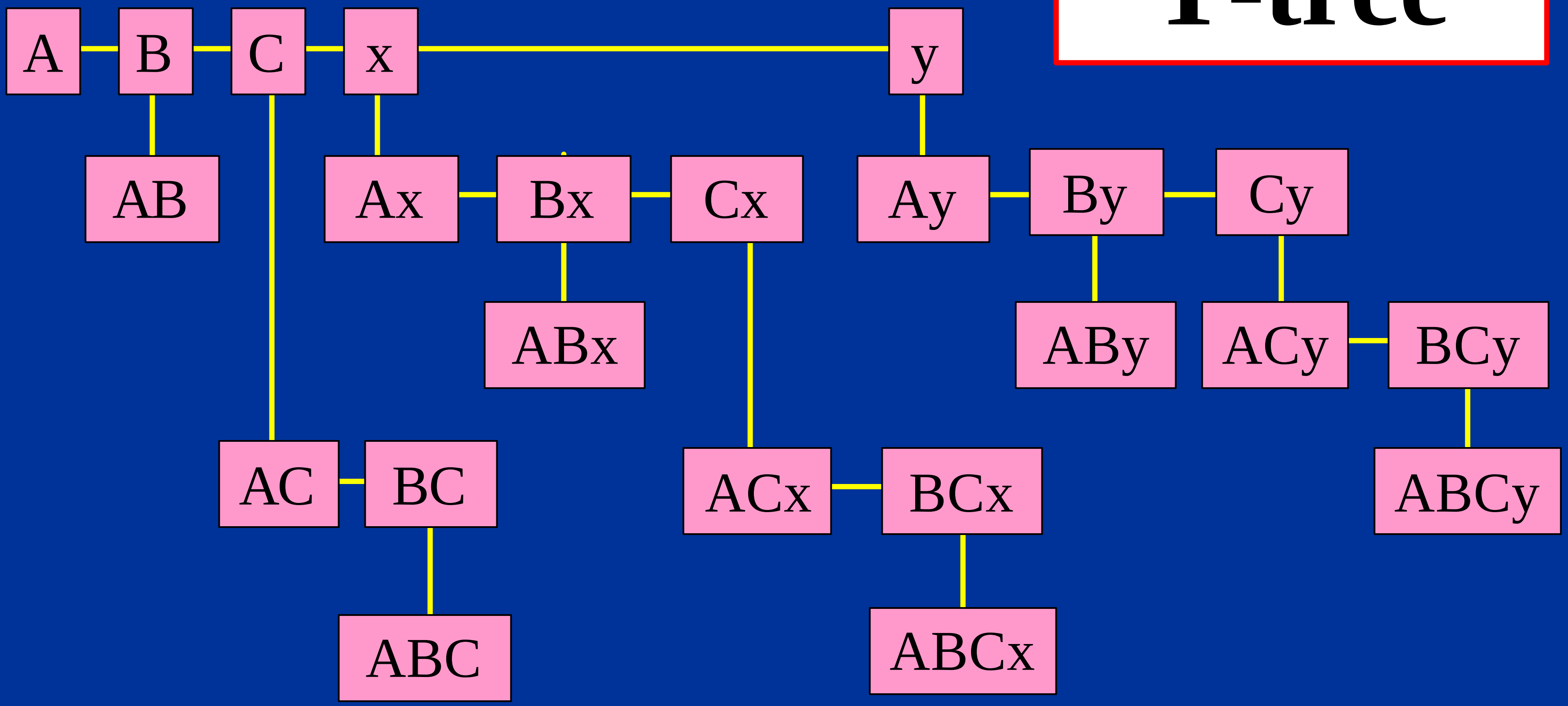
Total From Partial (TFP)

TFPC is an extension of the TFP Association Rule Mining (ARM) algorithm, and operates as follows:

1. Process data and store in a P-tree (Partial support tree), a set enumeration tree style structure which, as a by-product of its generation, includes at least partial counts for all relevant itemsets.
2. Generate a T-tree (Total support tree) from the P-tree using an Apriori style approach.
3. On completion the T-tree will comprise all the frequent item sets present in the data set together with their support counts.

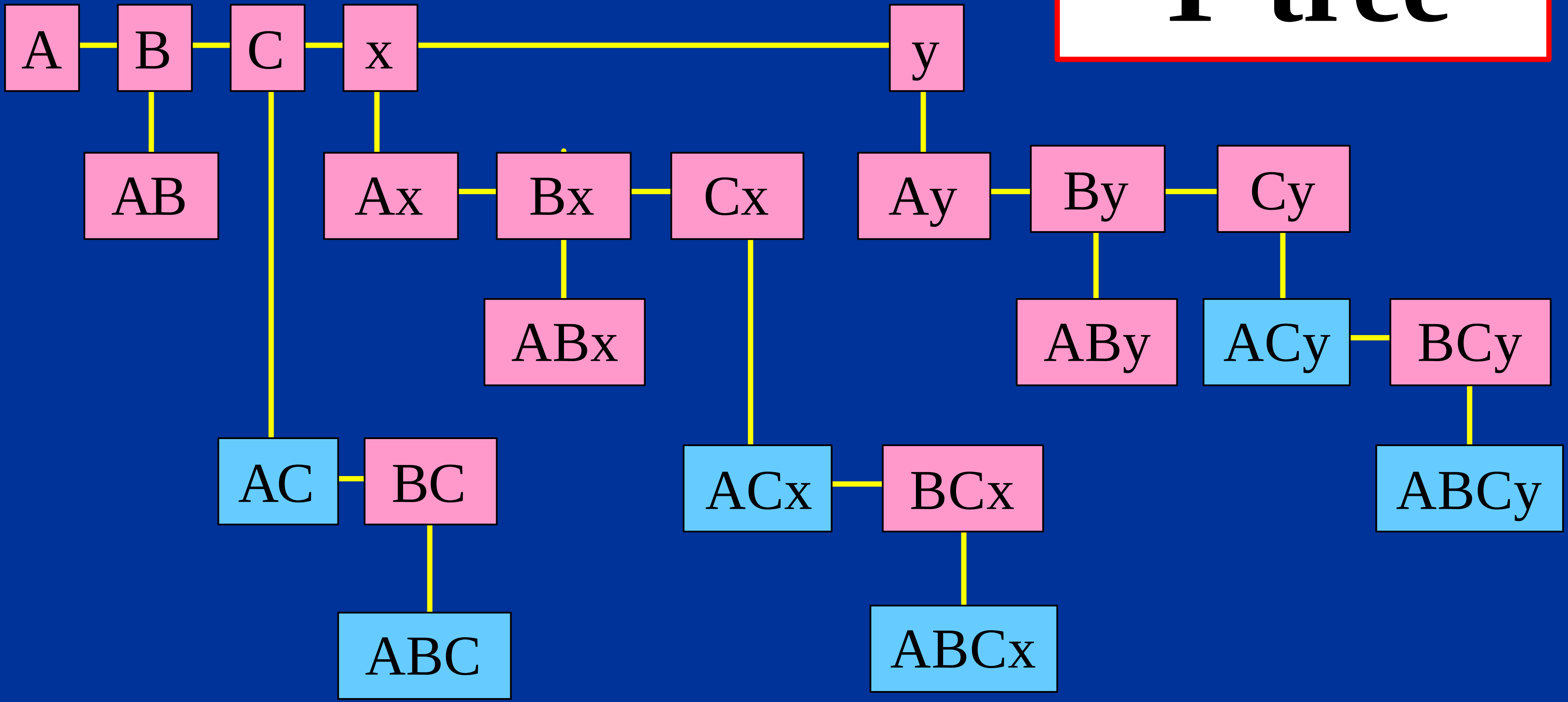
We have adapted TFP to generate CARs.

T-tree



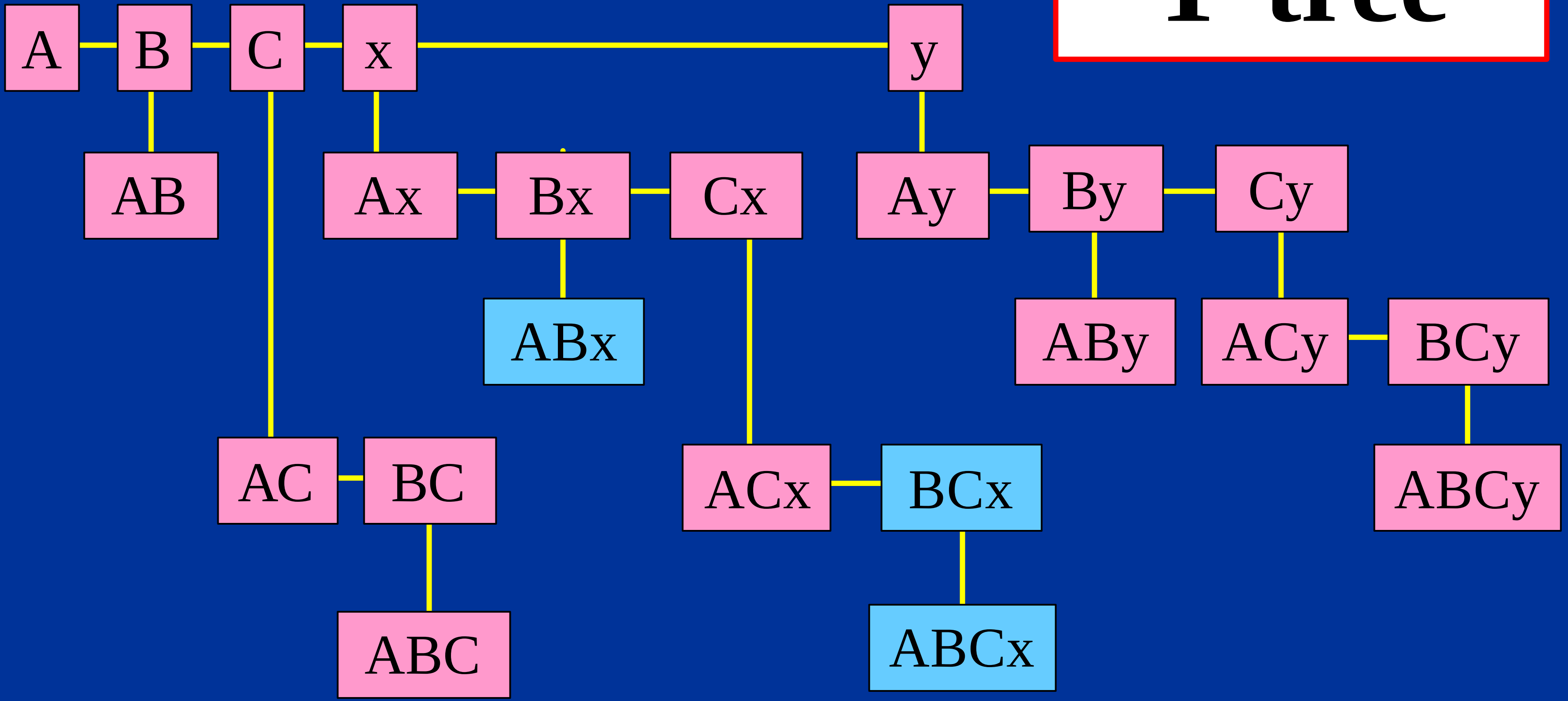
- T-tree is complete (except that combinations including both x and y do not appear).
- All itemsets that include x (or y) are in the sub-tree rooted in x (or y).

T-tree



- If AC not supported, then
- No candidates supersets of AC generated.

T-tree



- If itemset Bx gives $B \Rightarrow x$ with confidence above the given threshold, then
- no candidates supersets of Bx generated.

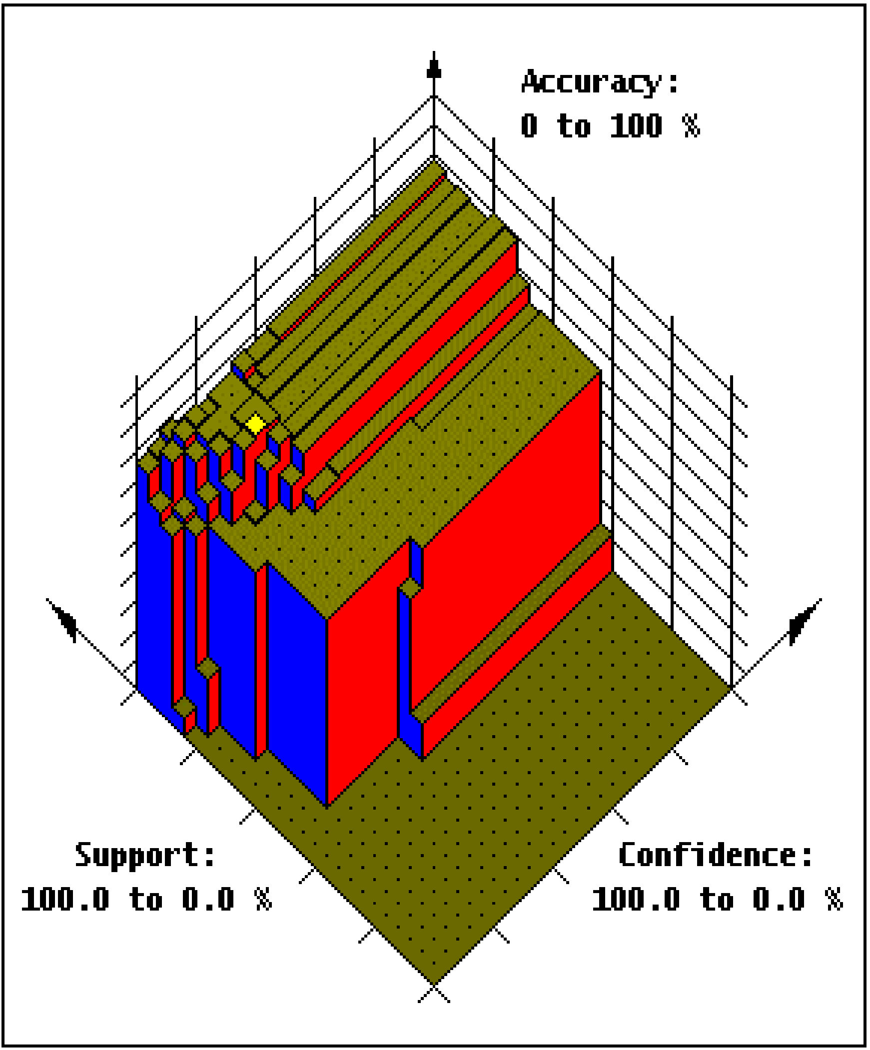
INITIAL RESULTS

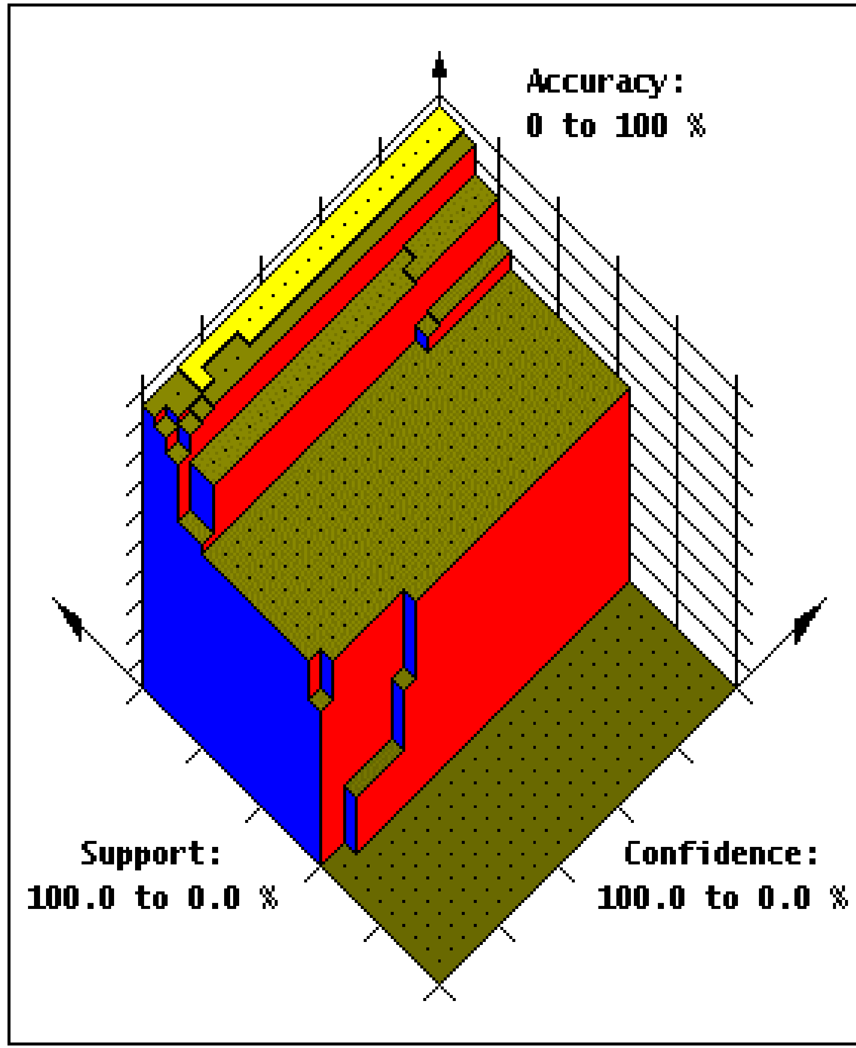
- Comparisons with CMAR (Li et al 2001) and CPAR (Yin and Han 2003).

Main findings:

1. TFPC significantly more efficient (ratio of 15:1 with respect to CMAR, and 46:1 with respect to CPAR).
2. Accuracy almost as good.

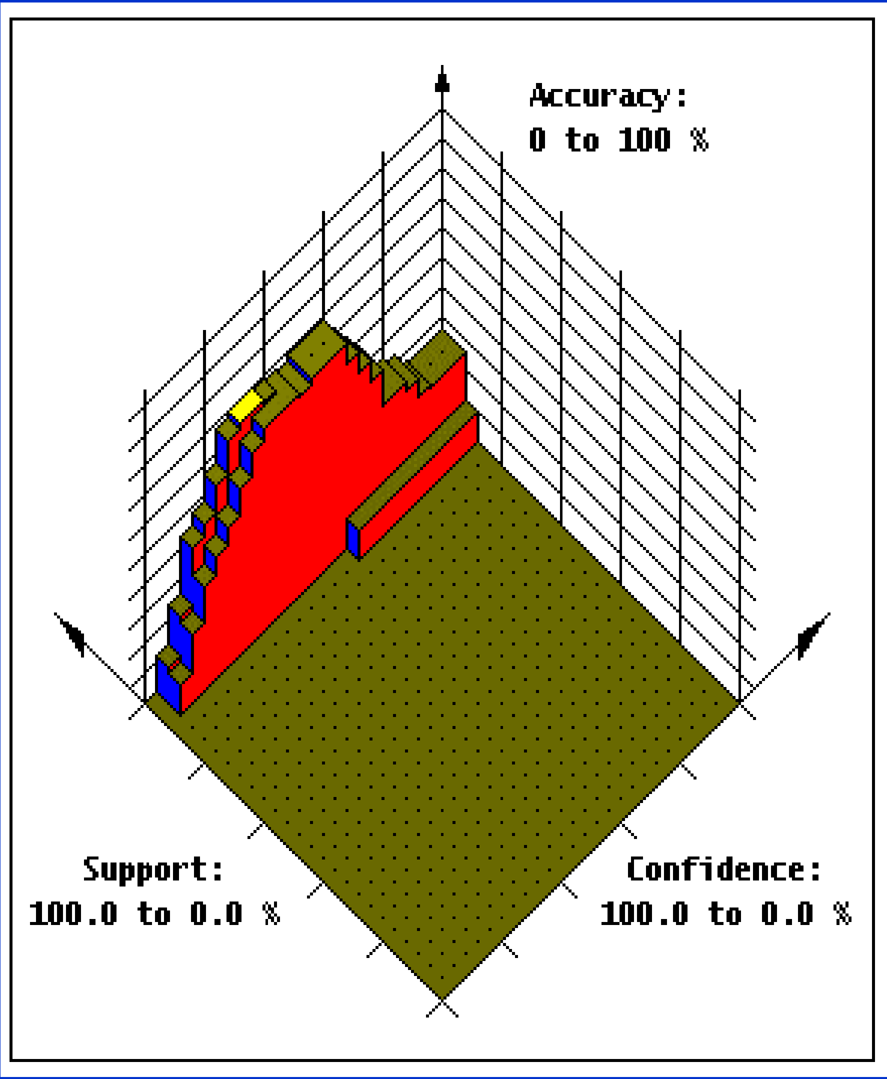
- Used 50% confidence and 1% support (as used by Li et al in their CMAR experiments).
- Could better results be obtained by tuning these thresholds?

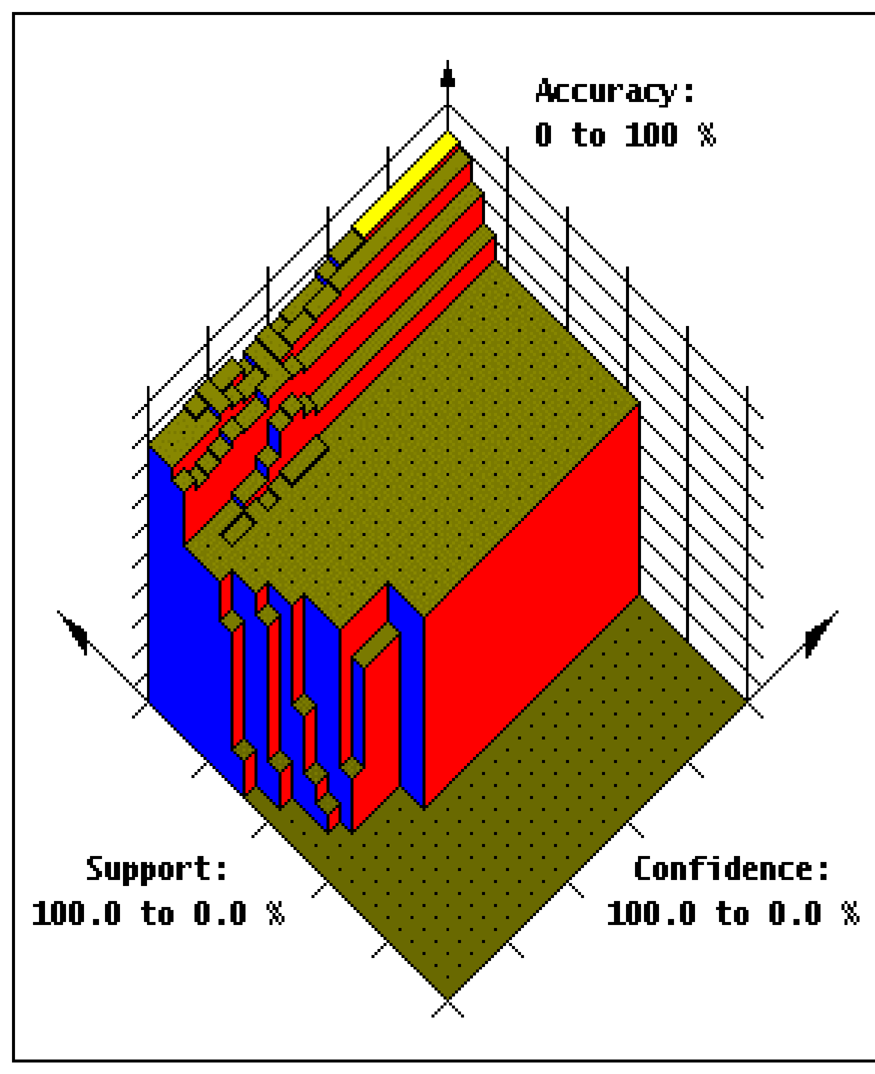




Horse Colic

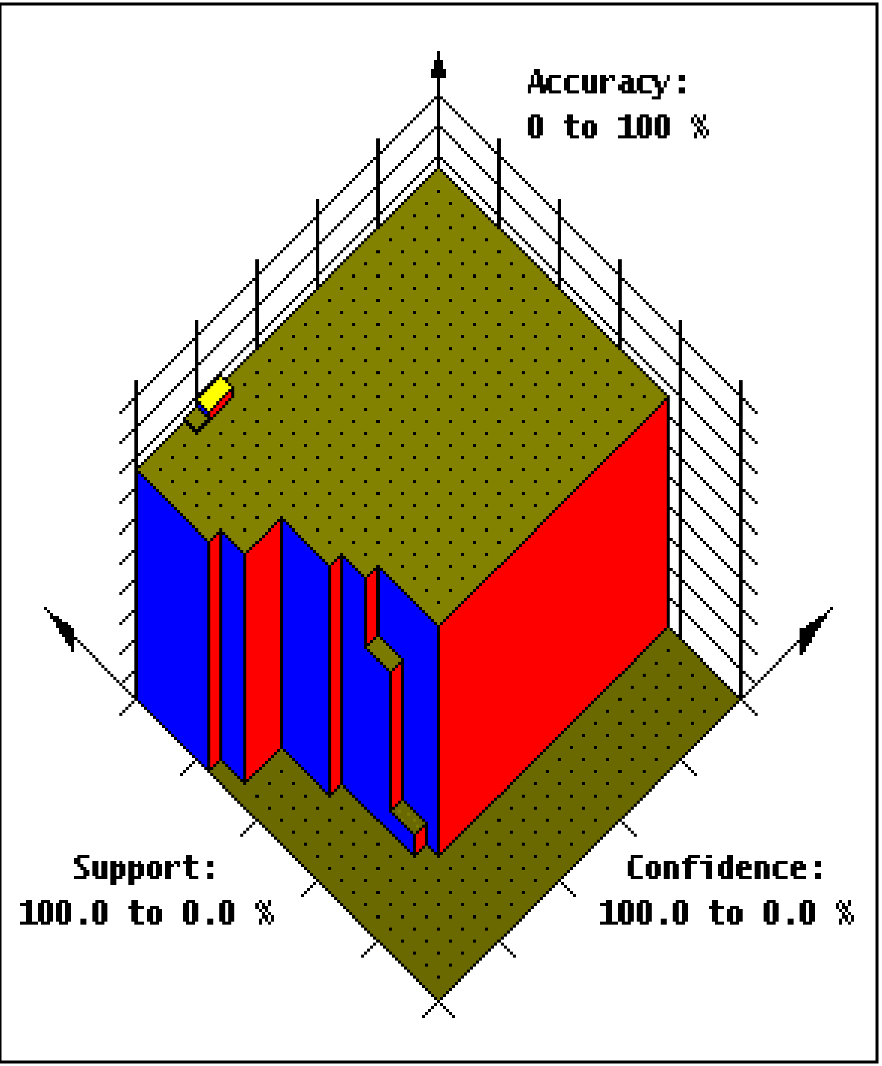
Breast Cancer

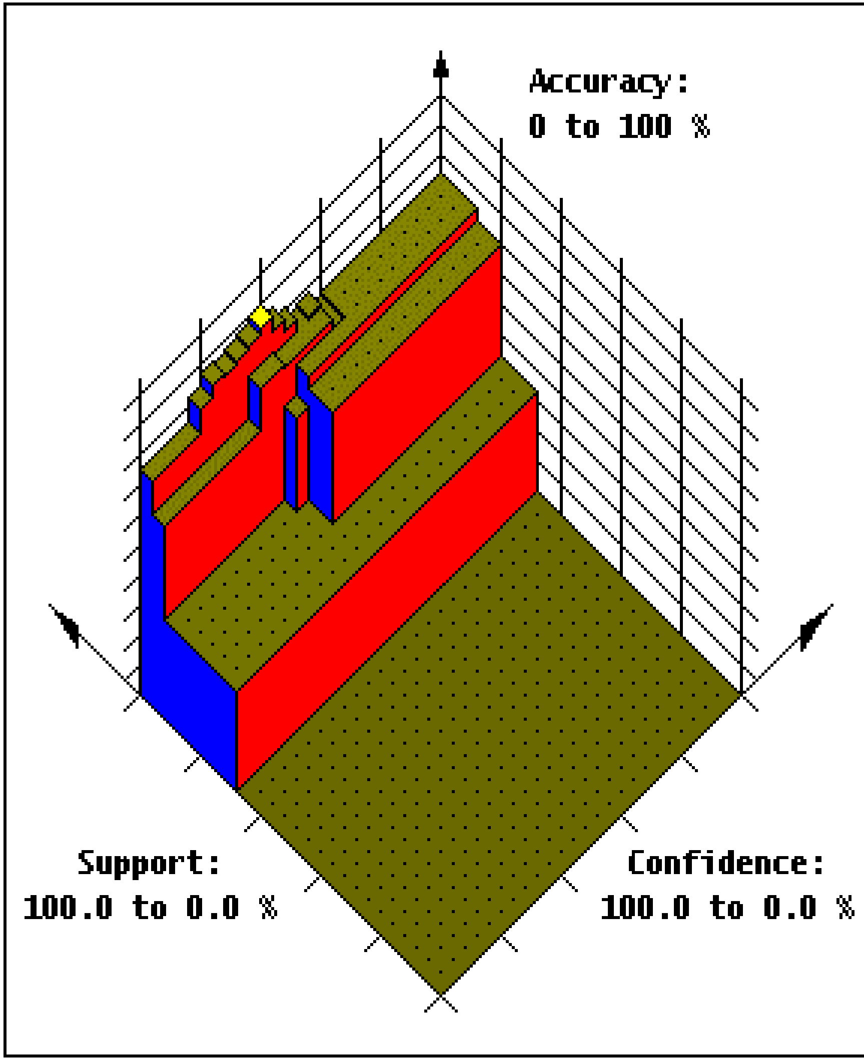




Led 7

Ionosphere





Adult

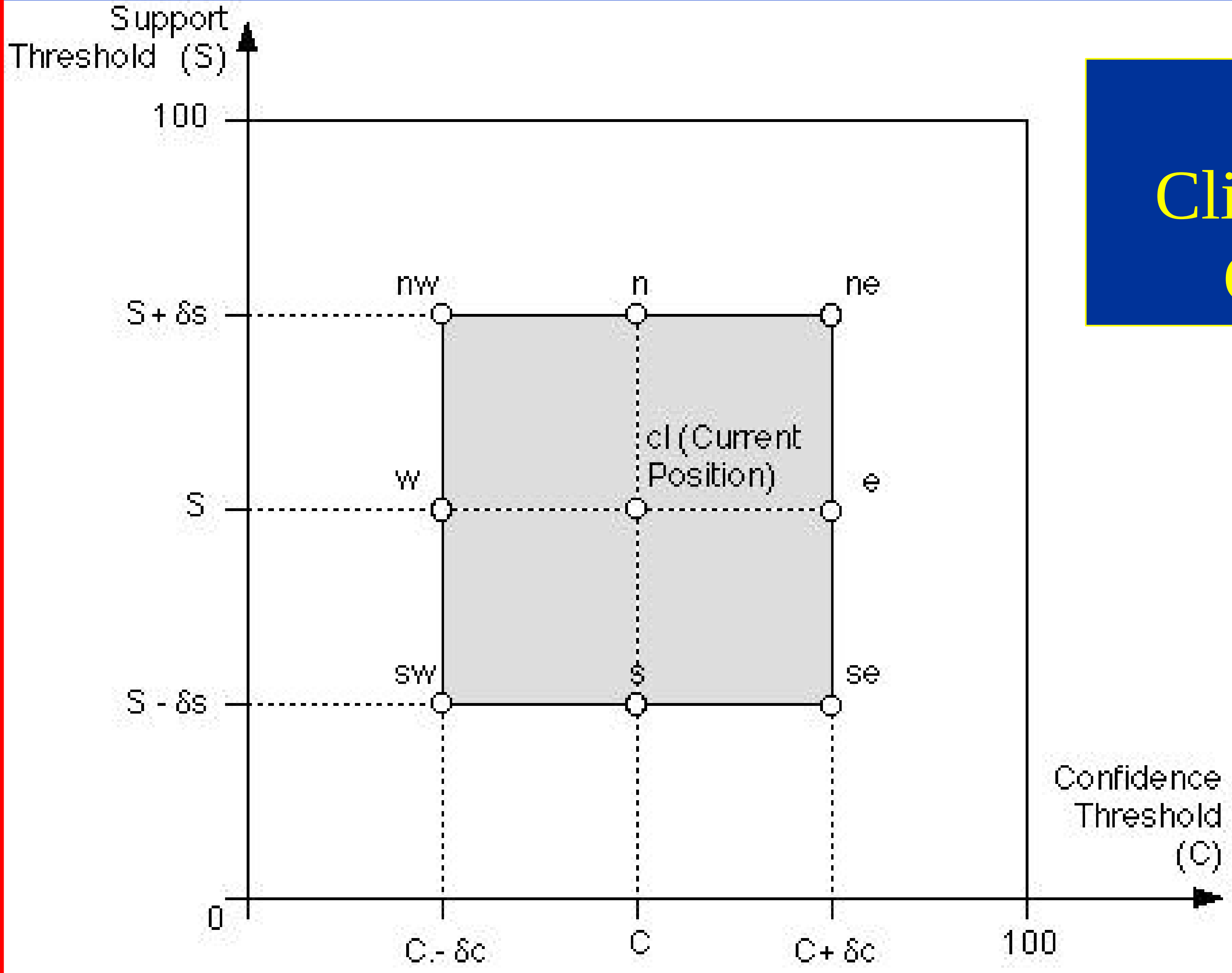
Nursery

FINDING BEST SUPPORT AND CONFIDENCE

- Apply TFPC algorithm in an iterative manner in conjunction with a “hill climbing” technique to traverse the 3-D “support-confidence-accuracy” space.
- Start with accuracy calculated for some start point.
- Calculate grid of points surrounding start point then either:

1. If centre has best accuracy “zero in” (reduce size of grid).
2. Else move to new location and repeat.

Hill Climbing Grid



δc = change in confidence threshold, δs = change in support threshold

RESULTS USING TFPC-HC

Result:

- TFPC with hill climbing (TFPC-HC) produces significantly better accuracy than CMAR and CBA, but
- At cost of computational efficiency (ratio of 3:1 compared to CMAR, but 1:1 compared to CPAR).

Refinement idea: Perform hill climbing only on first 9/10th to establish a “best” confidence and support threshold and then used this on the remaining 9/10ths.

RESULTS USING TFPC-HC+

Results:

- TFPC-HC+ still produces good accuracy (generally better than CMAR and CPAR).
- Computational efficiency advantages regained (ratio of 1:3 with respect to CMAR and 1:9 with respect to CPAR).

CONCLUSION AND FURTHER WORK

- TFPC-HC+ is a computationally efficient approach to tune confidence/support thresholds with a view to maximising CAR accuracy.
- The study presented here is limited (could consider other approaches, e.g. RIPPER).
- Would similar advantages be gained if HC was applied to (say) CMAR or CBA?