FACULTY
OF INFORMATICS
Masaryk University

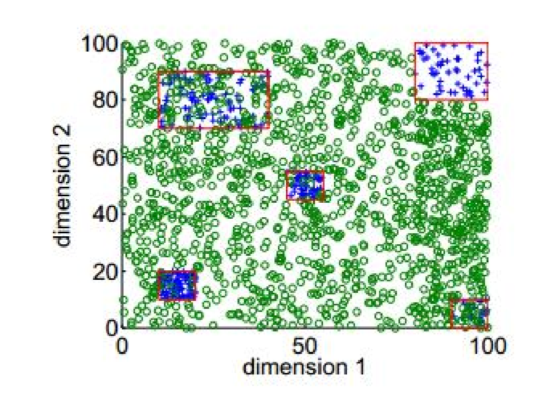
# Box drawings

## for learning with imbalanced data

KDD 2014

# Think of...

A few positive examples in a sea of negative examples. For modeling rare events: Machine breakdown, etc.



A box drawing classifier is a union of axis-parallel rectangles.

# The usual way to do this...

*Greedy*

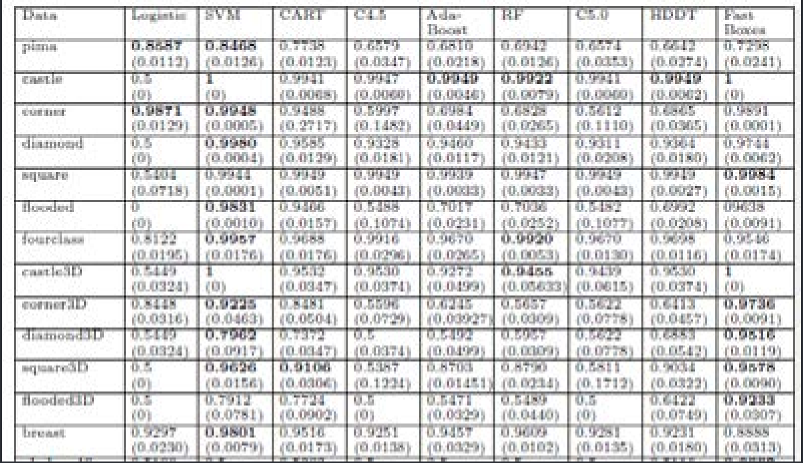
- Decision Tree (CART, C4.5 – Breiman 1984, Quinlan 1993)
  - Top down greedy: pick a features based on Gini Index or Information Gain, split data into two pieces, repeat. Prune afterwards.
- PRIM (Friedman, Fisher 1999)
  - Peel off subsets of data greedily, and if there is improvement, keep peeling off data. Occasionally put the data back.

This is too greedy for us

# Better solutions

- Approach 1: The Exact Boxes algorithm
  - Optimize weighted accuracy, regularize by number of boxes
  - Mixed-Integer Programming formulation
  - Useful for not-huge problems, but solves exactly the problem we care about
  - Acts as a gold standard to compare with because it solves exactly the problem we want.
- Approach 2:The Fast Boxes algorithm
  - Approximates the solution of Exact Boxes
  - Characterize (one class learning) before discriminating
  - Requires that features are continuous.

# Experiments



| Data | Logistic | SVM | CART | C4.5 | Ada-Boost | RF | C5.0 | HDDT | Fast Boxes |
|---|---|---|---|---|---|---|---|---|---|
| pima | **0.8587**<br>(0.0112) | **0.8468**<br>(0.0126) | 0.7738<br>(0.0123) | 0.6579<br>(0.0347) | 0.6810<br>(0.0218) | 0.6942<br>(0.0126) | 0.6574<br>(0.0353) | 0.6642<br>(0.0274) | 0.7298<br>(0.0241) |
| castle | 0.5<br>(0) | **1**<br>(0) | 0.9941<br>(0.0068) | 0.9947<br>(0.0069) | **0.9949**<br>(0.0046) | **0.9922**<br>(0.0079) | 0.9941<br>(0.0060) | **0.9949**<br>(0.0062) | **1**<br>(0) |
| corner | **0.9871**<br>(0.0129) | **0.9948**<br>(0.0005) | 0.9488<br>(0.2717) | 0.5997<br>(0.1482) | 0.6984<br>(0.0449) | 0.6828<br>(0.0265) | 0.5612<br>(0.1110) | 0.6865<br>(0.0365) | 0.9891<br>(0.0001) |
| diamond | 0.5<br>(0) | **0.9980**<br>(0.0004) | 0.9585<br>(0.0129) | 0.9328<br>(0.0181) | 0.9460<br>(0.0117) | 0.9433<br>(0.0121) | 0.9311<br>(0.0208) | 0.9364<br>(0.0180) | 0.9744<br>(0.0062) |
| square | 0.5404<br>(0.0718) | 0.9944<br>(0.0001) | 0.9949<br>(0.0051) | 0.9949<br>(0.0043) | 0.9939<br>(0.0033) | 0.9947<br>(0.0032) | 0.9949<br>(0.0043) | 0.9949<br>(0.0027) | **0.9984**<br>(0.0015) |
| flooded | 0<br>(0) | **0.9831**<br>(0.0010) | 0.9406<br>(0.0157) | 0.5388<br>(0.1074) | 0.7017<br>(0.0231) | 0.7036<br>(0.0252) | 0.5482<br>(0.1077) | 0.6992<br>(0.0208) | 09638<br>(0.0091) |
| fourclass | 0.8122<br>(0.0195) | **0.9957**<br>(0.0176) | 0.9688<br>(0.0176) | 0.9916<br>(0.0296) | 0.9670<br>(0.0265) | **0.9920**<br>(0.0053) | 0.9670<br>(0.0130) | 0.9698<br>(0.0116) | 0.9546<br>(0.0174) |
| castle3D | 0.5449<br>(0.0324) | **1**<br>(0) | 0.9532<br>(0.0347) | 0.9530<br>(0.0374) | 0.9272<br>(0.0499) | **0.9455**<br>(0.05633) | 0.9439<br>(0.0615) | 0.9530<br>(0.0374) | **1**<br>(0) |
| corner3D | 0.8448<br>(0.0316) | **0.9225**<br>(0.0463) | 0.8481<br>(0.0504) | 0.5596<br>(0.0729) | 0.6245<br>(0.03927) | 0.5657<br>(0.0309) | 0.5622<br>(0.0778) | 0.6413<br>(0.0457) | **0.9736**<br>(0.0091) |
| diamond3D | 0.5449<br>(0.0324) | **0.7962**<br>(0.0917) | 0.7372<br>(0.0347) | 0.5<br>(0.0374) | 0.5492<br>(0.0499) | 0.5967<br>(0.0309) | 0.5622<br>(0.0778) | 0.6883<br>(0.0542) | **0.9516**<br>(0.0119) |
| square3D | 0.5<br>(0) | **0.9626**<br>(0.0156) | **0.9106**<br>(0.0306) | 0.5387<br>(0.1224) | 0.8703<br>(0.01451) | 0.8790<br>(0.0234) | 0.5811<br>(0.1712) | 0.9034<br>(0.0322) | **0.9578**<br>(0.0090) |
| flooded3D | 0.5<br>(0) | 0.7912<br>(0.0781) | 0.7724<br>(0.0902) | 0.5<br>(0) | 0.5471<br>(0.0329) | 0.5489<br>(0.0440) | 0.5<br>(0) | 0.6422<br>(0.0749) | **0.9233**<br>(0.0307) |
| breast | 0.9297<br>(0.0230) | **0.9801**<br>(0.0079) | 0.9516<br>(0.0173) | 0.9251<br>(0.0138) | 0.9457<br>(0.0329) | 0.9609<br>(0.0102) | 0.9281<br>(0.0135) | 0.9231<br>(0.0180) | 0.8888<br>(0.0313) |

# Summary

- Exact Boxes
  - Mixed integer Programming
- Fast Boxes
  - Characterize-then-discriminate approach.
- Take away: Aim for both interpretability and accuracy, because you can often get both.

Thank You!