Data Mining for Dataset #3

Results and Discussion

Henry Xiao

xiao@cs.queensu.ca

School of Computing

Queen's University

Preliminary Review

from last week's preliminary studies on this dataset:

- 10 attributes and 3 classes. (2-class case is trivial.)
- Class 1 can be clearly discriminated from Class 2 and 3.
- Class 2 is difficult to be separated from Class 3.
- Directly applying 3 methods results:

Method	BayesNet	DecisionTable	PRISM
Correctness	95.2496	98.9521	98.7426

Current Process

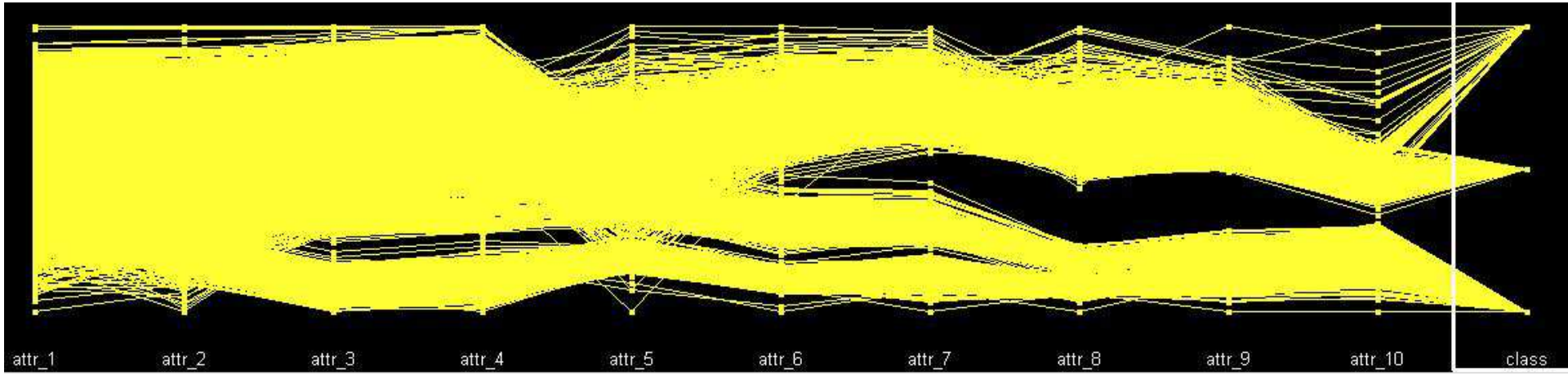
Consider following perspectives:

- Attribute selection - attribute subset.
- Exploration from visualization.
- Method selection - *BayesNet*, *DecisionTable*, and *PRISM*.
- Possibility of existing other classes from the SVD/SDD plot.

Queen's UNIVERSITY

3-Class Visualization

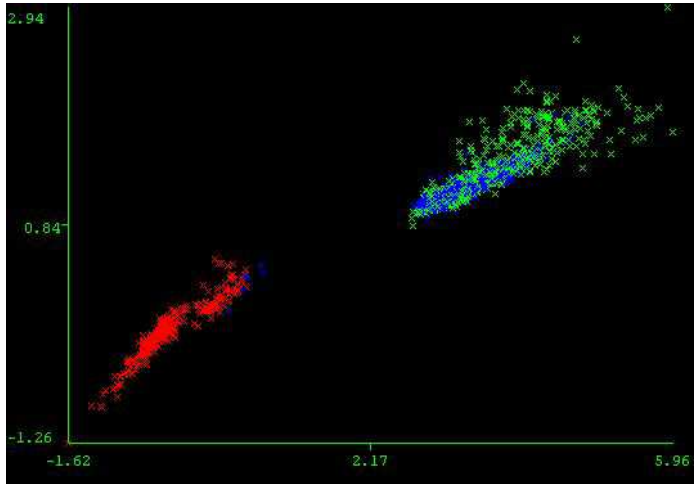
3-class parallel visualization:



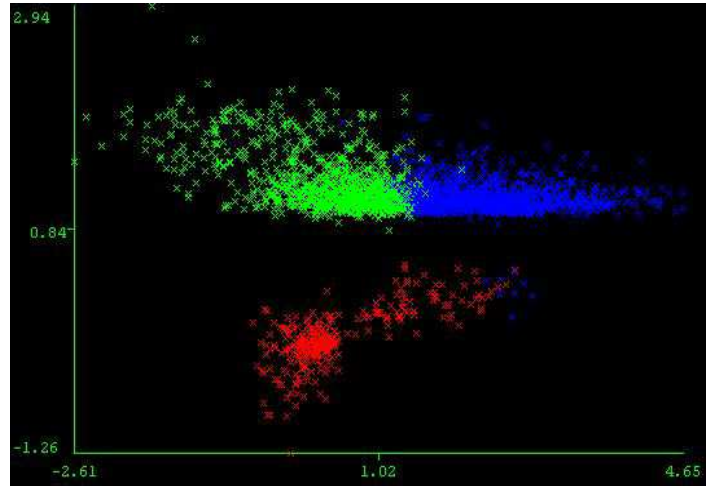
Powerful plot here!

Cluster Visualization

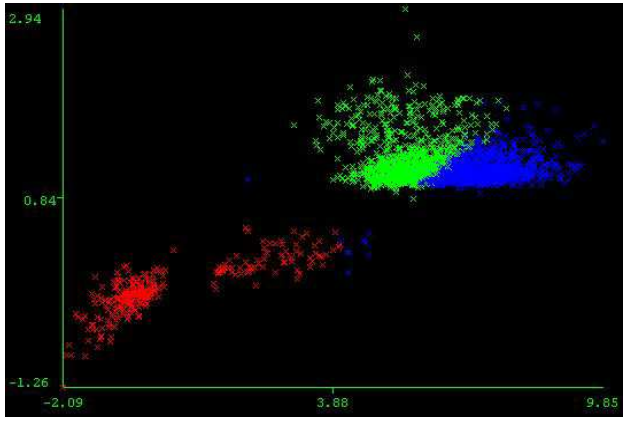
Different cluster graphs (k-Mean) are demonstrated below:



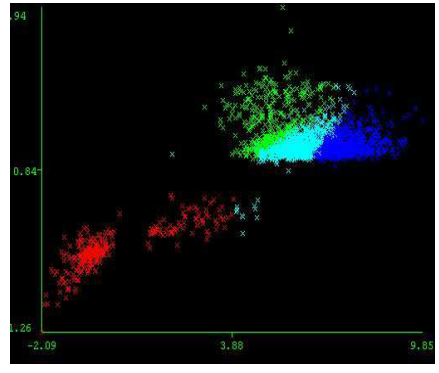
Attribute9 - 10



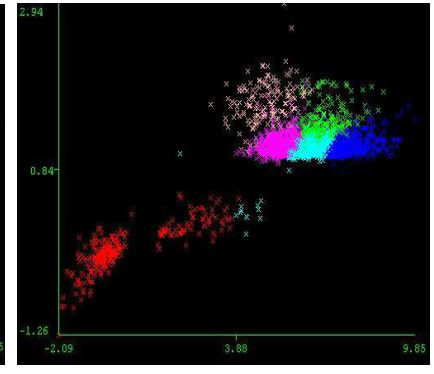
Attribute1 -10



Attribute4-10



4 - class



6 - class

Explore Visualization

Possible observations from the graphs:

- Class 1 can be easily identified from clustering.
- Two classes (2 and 3) are joint and overlapping.
- Class 2 and 3 are hard to be discriminated.
- The mass containing Class 2 and 3 may be further categorized.
- Class 1 may also be categorized into two well separated classes.
- 3 classes may not be enough to capture the attributes of this dataset.

Queen's UNIVERSITY

Attribute Selection

Attribute subsets are selected from different ways.

- $\{4, 10\}$ - from attribute visualization.
- $\{1, 4, 10\}$ - attribute 1 seems helpful.
- $\{1, 4, 7, 10\}$ - DecisionTable attribute subset.
- $\{4, 9, 10\}$ - from InfoGain ranking.

Queen's UNIVERSITY

Preliminary Result

Some results are shown in the table.

Attribute Set	*BayesNet*%	*DecisionTable*%	*PRISM*%
$\{4, 10\}$	98.5153	98.7647	96.9355
$\{1, 4, 10\}$	98.5271	98.7647	97.6244
$\{1, 4, 7, 10\}$	98.2658	98.9429	98.2896
$\{4, 9, 10\}$	98.5984	98.7766	97.5056
whole set	95.2251	98.8241	98.4915

All tests are done with cross validation.

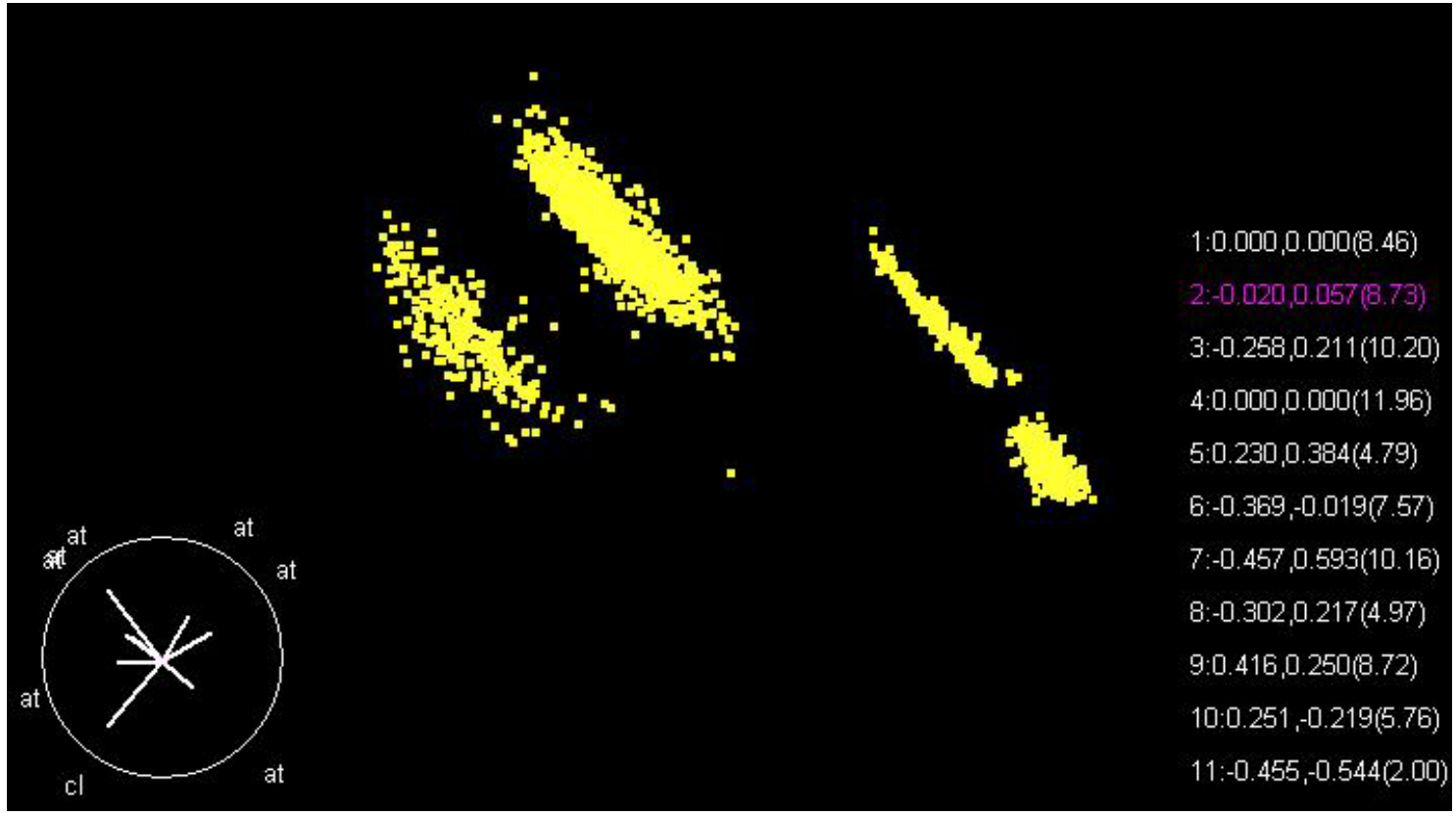
Result Cont's

Remarks on the results.

- Attribute set $\{1, 4, 7, 10\}$ looks good.
- Three methods results are comparable.
- DecisionTable is slightly better than PRISM.
- Rule base methods seems to make more sense.
- Nice dataset property of only 10 attributes.

Discussion

What can be done to get a better classification?



Scatter Plot.

Ending

Questions regarding mining results?

Information Site: http://www.cs.queensu.ca/home/xiao/dm.html

E-mail: xiao@cs.queens.ca

Thank you