Data Mining for Additional Dataset

Not Bored But Curious

Henry Xiao

xiao@cs.queensu.ca

School of Computing

Queen's University

Queen's UNIVERSITY

Dataset Preview

This dataset contains a significant amount of missing values.

- Attributes and Samples are the same as the previous dataset.
- Main properties still hold in this dataset.
- Main problem is still to identify “Red”(i.e.,7) and “Brown”(i.e.,4).
- 8419 samples in total:
  - 256 “Brown” - 4
  - 7186 “Red” - 7
  - 977 “White” - 8

Queen's UNIVERSITY

Deal Missing Value

We consider two ways to deal with the missing values:

- Intuitive method - Ignore the missing values.
  - No preprocessing needed.
  - *Weka* directly supports this method.
  - Counting less samples for each attribute.
- Replace method - Fill the missing values with the respected means.
  - Calculating each attribute mean with each class.
  - Replacing the missing value with “the” mean.
  - *MatLab* code can be programmed to preprocess the dataset.
  - Counting all samples for each attribute.

Queen's UNIVERSITY

The Means

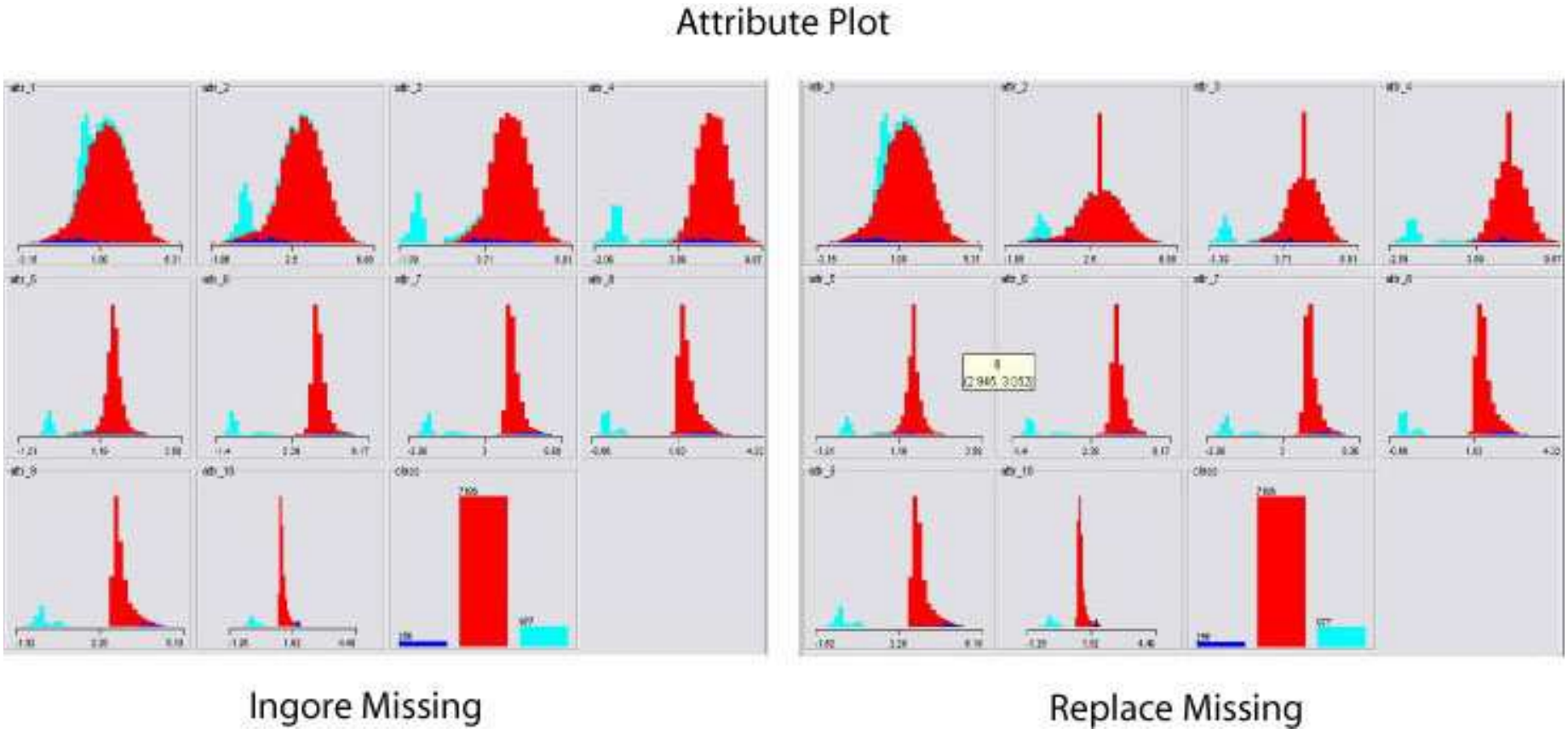
Following table lists all the means respecting to three classes.

Attribute	*1*	*2*	*3*	*4*	*5*
4	-0.098	1.317	4.051	5.958	1.475
7	1.498	3.081	5.198	6.370	1.594
8	0.559	0.478	0.325	0.150	-0.060
Attribute	*6*	*7*	*8*	*9*	*10*
4	4.147	6.047	2.657	4.542	1.888
7	3.701	4.871	2.108	3.277	1.170
8	-0.200	-0.367	-0.145	-0.314	-0.169

Queen's UNIVERSITY

Attribute Plot

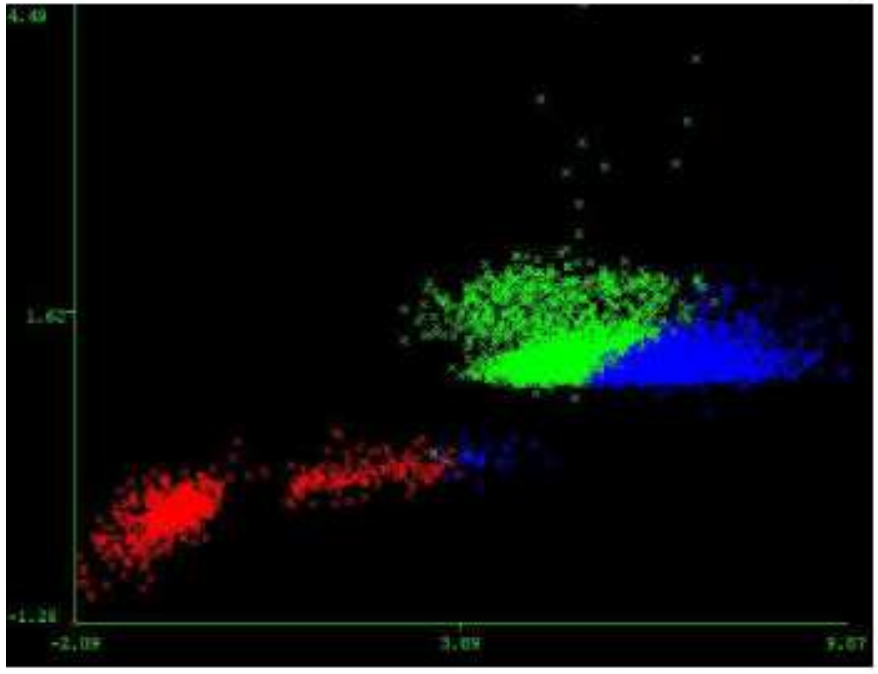
Attribute Plot for the two processed datasets with the two methods.



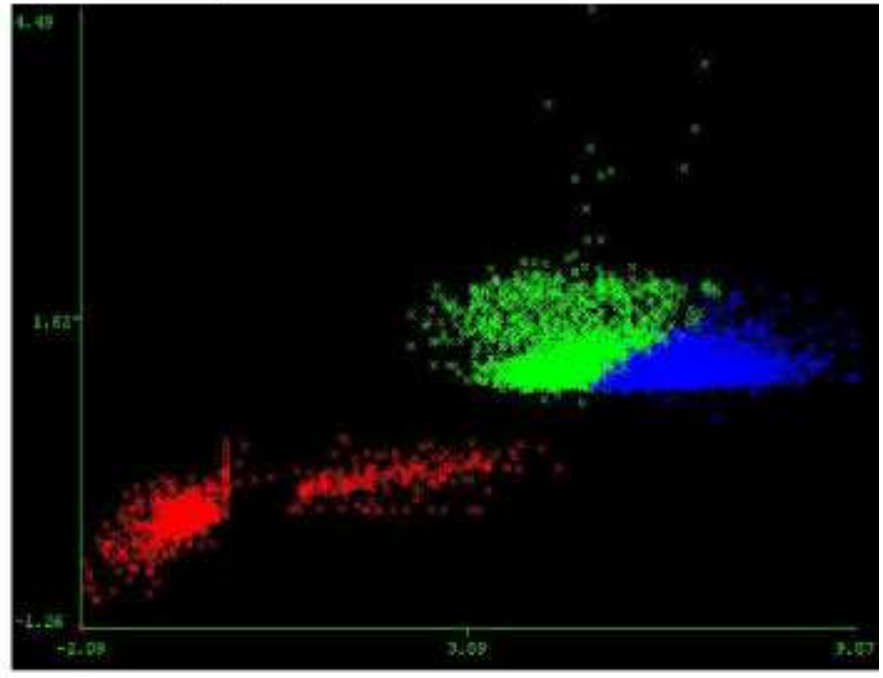
Cluster Visualization

3 cluster visualizations for both methods.

Cluster (Attribute 4 - 10)



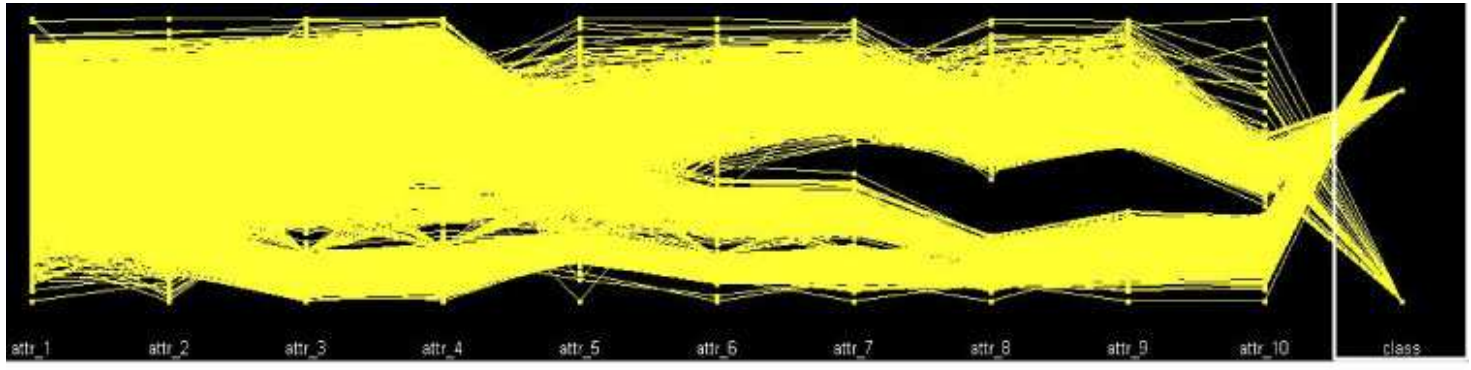
Ingore Missing



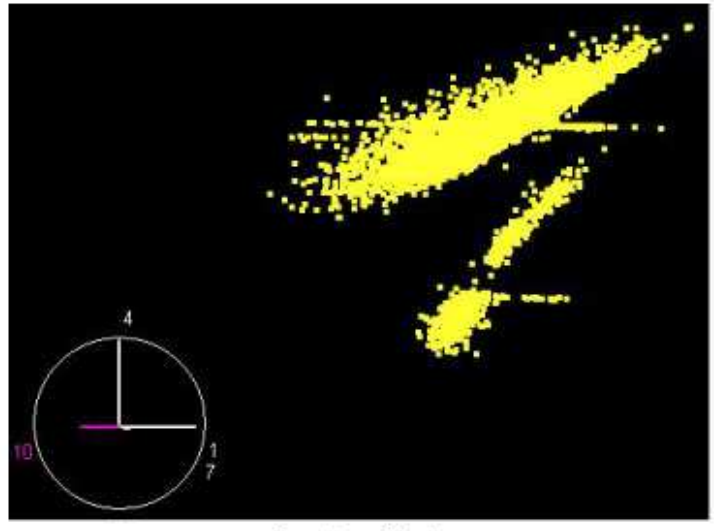
Replace Missing

Other Visualization

Try to explore the effect after replacing the missing values.



Parallel Plot



Scatter Plot

Ming Processed Dataset

Some results are shown in the table.

Dataset	Attribute Set	*BayesNet*%	*DecisionTable*%	*PRISM*%
Ignore	$\{10\}$	98.4915	98.4915	n/a
Replace	$\{10\}$	98.7885	98.7885	14.6454
Ignore	$\{1, 4, 7, 10\}$	98.0639	98.6459	n/a
Replace	$\{1, 4, 7, 10\}$	98.1946	98.9072	54.0207
Ignore	whole set	95.2013	98.6697	n/a
Replace	whole set	95.8071	98.9072*	93.9898

DecisionTable uses a subset $\{6, 7, 9, 10\}$.

Queen's UNIVERSITY

Result Observations

Some observations from our experiments.

- Sorting affects PRISM significantly.
- Replacing by mean helps the correlation.
- Replacing by mean doesnot change the information gain ranking, but decreases the ratio.
- Replacing policy is more helpful statistically.
- Confusion Matrix (PRISM, $\{10\}$): $\left(\begin{array}{ccc|c} 256 & 0 & 0 & 4 \\ 7186 & 0 & 0 & 7 \\ 0 & 0 & 977 & 8 \end{array}\right)$

Discussion

- What is the practical meaning of the missing values?
- What is a proper method to deal with the missing values?
- What is the effect after processing the dataset?
- What is the effect towards different mining techniques?

Queen's UNIVERSITY

Ending

Questions regarding mining results?

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

E-mail: xiao@cs.queens.ca

Thank you