Analysis of Dataset ♯1

Preliminary Report

Henry Xiao

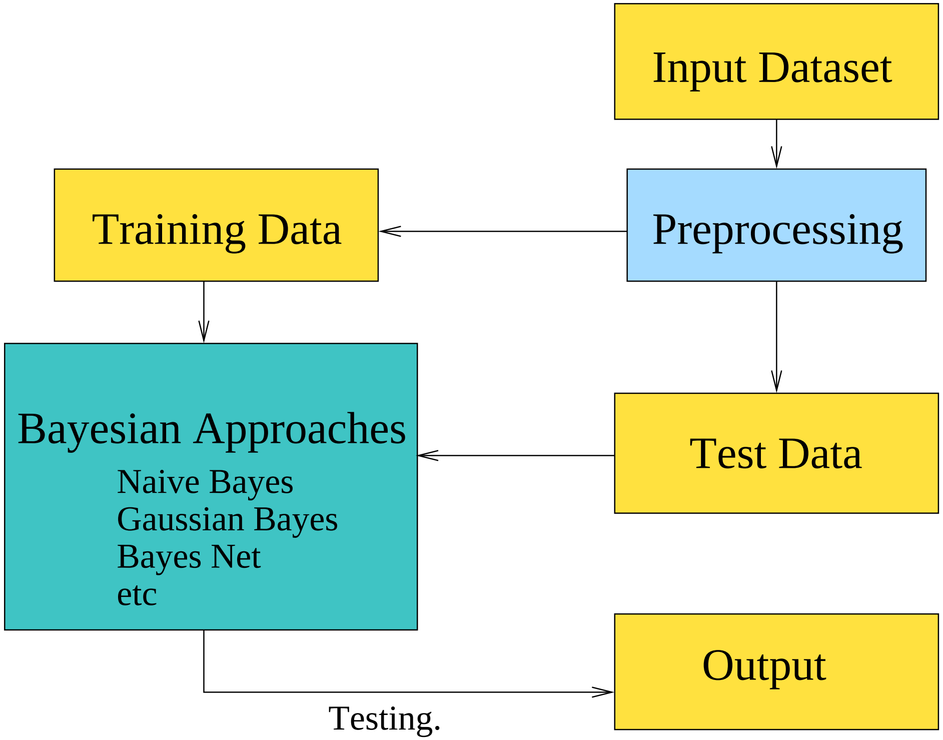
xiao@cs.queensu.ca

School of Computing

Queen’s University

Testing Model

We use *Weka* to analysis the dataset with different Bayesian Approaches. The basic model is showed in the below figure.



Dataset Bases

Weka Explorer gives basic information of our dataset once loaded in.

Dataset Bases

Weka Explorer gives basic information of our dataset once loaded in.

- 138 instances and 136 attributes.

Dataset Bases

Weka Explorer gives basic information of our dataset once loaded in.

- 138 instances and 136 attributes.
- Class Variable: $Target \in 0, 1$.

Dataset Bases

Weka Explorer gives basic information of our dataset once loaded in.

- 138 instances and 136 attributes.
- Class Variable: $Target \in 0, 1$.
- Six attributes are clearly just labels or counters $(Main\_Case, SAMPLE, AMGE, AMGN, Class, Regolith)$.

Dataset Bases

Weka Explorer gives basic information of our dataset once loaded in.

- 138 instances and 136 attributes.
- Class Variable: $Target \in 0, 1$.
- Six attributes are clearly just labels or counters $(Main\_Case, SAMPLE, AMGE, AMGN, Class, Regolith)$.
- All attributes are numeric, which need to be discretized into nominal.

Discretize and Attribute Selection

We need to discretize the attributes to nominal in order to use any Bayesian approach. Intuitively, we also need to select attributes for mining.

Discretize and Attribute Selection

We need to discretize the attributes to nominal in order to use any Bayesian approach. Intuitively, we also need to select attributes for mining.

- *Weka* provides a class to discretize the attributes (`weka.filters.unsupervised.attribute`).

Discretize and Attribute Selection

We need to discretize the attributes to nominal in order to use any Bayesian approach. Intuitively, we also need to select attributes for mining.

- *Weka* provides a class to discretize the attributes (`weka.filters.unsupervised.attribute`).
- Two things need to be chosen for the attribute selection (`weka.filters.supervised.attribute.AttributeSelecti`

Discretize and Attribute Selection

We need to discretize the attributes to nominal in order to use any Bayesian approach. Intuitively, we also need to select attributes for mining.

- *Weka* provides a class to discretize the attributes (`weka.filters.unsupervised.attribute`).
- Two things need to be chosen for the attribute selection (`weka.filters.supervised.attribute.AttributeSelecti`
  - *Evaluator*: Determines how attributes/attribute subsets are evaluated.
    - *CfsSubsetEval*: CFS attribute subset evaluator.
    - *ClassifierSubsetEval*: Classifier subset evaluator.
    - *InfoGainAttributeEval*: Evaluating attributes individually by measuring the information gain.

Discretize and Attribute Selection

We need to discretize the attributes to nominal in order to use any Bayesian approach. Intuitively, we also need to select attributes for mining.

- *Weka* provides a class to discretize the attributes (`weka.filters.unsupervised.attribute`).
- Two things need to be chosen for the attribute selection (`weka.filters.supervised.attribute.AttributeSelecti`
  - *Search*: Determines the search method.
    - *BestFirst, GreedyStepwise, RandomSearch, ExhaustiveSearch* (intuitive).
    - *RaceSearch, Ranker, RankSearch* (unknown).

Selected Attributes

We apply different Evaluator and Search Method to get different attribute subset.

- *CfsSubsetEval* and *BestFirst*: 15 Attributes.
- *InfoGainAttributeEval* and *Ranker*: 12 Attributes (above 0.1000).
- *ClassifierSubsetEval-NaiveBayes* and *BestFirst*: 11 Attributes.

Detail information of the attribute subsets can be found on my page: *http://www.cs.queensu.ca/home/xiao/dm.html*.

Bayes Classifiers and Bayes Net

Review of available Bayes approaches at *Weka*.

Bayes Classifiers and Bayes Net

Review of available Bayes approaches at *Weka*.

- *NaiveBayes*: Estimate Posterior - $Y^{predict} = arg\max_v P(Y = v|X_1 = u_1, \ldots, X_m = u_m)$.

Bayes Classifiers and Bayes Net

Review of available Bayes approaches at *Weka*.

- *NaiveBayes*: Estimate Posterior - $Y^{predict} = arg\max_v P(Y = v|X_1 = u_1, \ldots, X_m = u_m)$.
- *NaiveBayesSimple*: Use a simple Naive Bayes classifier modelled by a normal distribution.

Bayes Classifiers and Bayes Net

Review of available Bayes approaches at *Weka*.

- *NaiveBayes*: Estimate Posterior - $Y^{predict} = arg\max_v P(Y = v | X_1 = u_1, \ldots, X_m = u_m)$.
- *NaiveBayesSimple*: Use a simple Naive Bayes classifier modelled by a normal distribution.
- *NaiveBayesUpdateable*: The updateable version of NaiveBayes.

Bayes Classifiers and Bayes Net

Review of available Bayes approaches at *Weka*.

- *NaiveBayes*: Estimate Posterior - $Y^{predict} = arg\max_v P(Y = v|X_1 = u_1, \ldots, X_m = u_m)$.
- *NaiveBayesSimple*: Use a simple Naive Bayes classifier modelled by a normal distribution.
- *NaiveBayesUpdateable*: The updateable version of NaiveBayes.
- *BayesNets*: Bayes Network learning.

Bayes Classifiers and Bayes Net

Review of available Bayes approaches at *Weka*.

- *NaiveBayes*: Estimate Posterior - $Y^{predict} = arg\max_v P(Y = v|X_1 = u_1, \ldots, X_m = u_m)$.
- *NaiveBayesSimple*: Use a simple Naive Bayes classifier modelled by a normal distribution.
- *NaiveBayesUpdateable*: The updateable version of NaiveBayes.
- *BayesNets*: Bayes Network learning.
- *AODE*: Achieve highly accurate classification by averaging over all of a small space of alternative naive-Bayes-like models that have weaker (and hence less detrimental) independence assumptions than naive Bayes.

Preliminary Results - All Attributes

We use all attributes (except removed 6) for the mining first.

Bayes Method	*Correct Rate%*	*Incorrect Rate%*	*Build Time(s)*
NaiveBayes	68.0851	31.9149	0.02
NaiveSimple	68.0851	31.9149	0.08
NaiveUpdate	68.0851	31.9149	0.02
BayesNets	65.9574	34.0426	0.03
AODE	n/a	n/a	0.39

The training and testing splitting is default at 66%.

Preliminary Results - 15 Attributes

We use selected attributes from the *CfsSubsetEval* and *BestFirst*.

Bayes Method	*Correct Rate%*	*Incorrect Rate%*	*Build Time(s)*
NaiveBayes	80.8511	19.1489	0
NaiveSimple	80.8511	19.1489	0
NaiveUpdate	80.8511	19.1489	0
BayesNets	80.8511	19.1489	0
AODE	78.7234	21.2766	0

The training and testing splitting is default at 66%.

Preliminary Results - 12 Attributes

We use selected attributes from the *InfoGainAttributeEval* and *Ranker*.

Bayes Method	*Correct Rate%*	*Incorrect Rate%*	*Build Time(s)*
NaiveBayes	80.8511	19.1489	0
NaiveSimple	80.8511	19.1489	0.02
NaiveUpdate	80.8511	19.1489	0
BayesNets	78.7234	21.2766	0
AODE	78.7234	21.2766	0.03

The training and testing splitting is default at 66%.

Preliminary Results - 11 Attributes

We use selected attributes from the *ClassifierSubsetEval-NaiveBayes* and *BestFirst*.

Bayes Method	*Correct Rate%*	*Incorrect Rate%*	*Build Time(s)*
NaiveBayes	80.8511	19.1489	0
NaiveSimple	80.8511	19.1489	0
NaiveUpdate	80.8511	19.1489	0
BayesNets	87.2340	12.7660	0.03
AODE	80.8511	19.1489	0

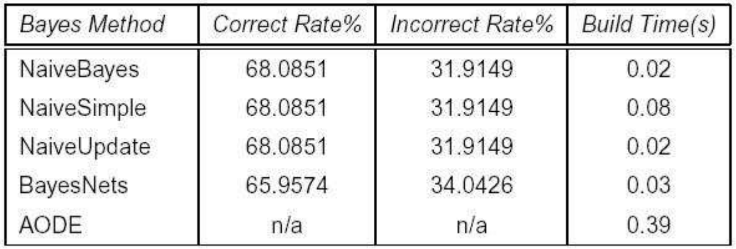
The training and testing splitting is default at 66%.

Discussion

What we get from the preliminary play around.

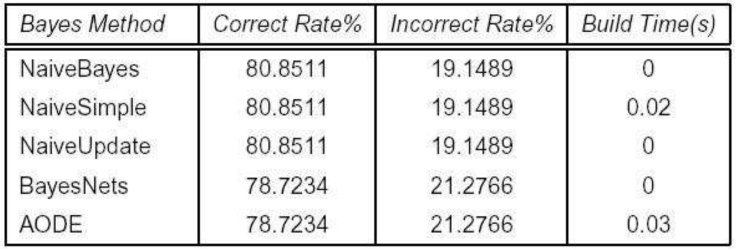
- The attribute selection is important.
- The performance of different Bayes method varies.
- Inference model requires more data preprocessing.
- Weka needs DOCUMENTATION!

All Attributes



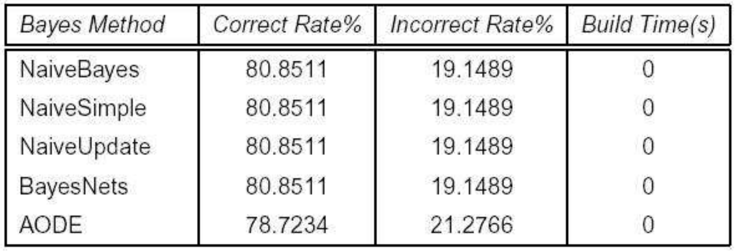
Bayes Method	Correct Rate%	Incorrect Rate%	Build Time(s)
NaiveBayes	68.0851	31.9149	0.02
NaiveSimple	68.0851	31.9149	0.08
NaiveUpdate	68.0851	31.9149	0.02
BayesNets	65.9574	34.0426	0.03
AODE	n/a	n/a	0.39

12 Attributes



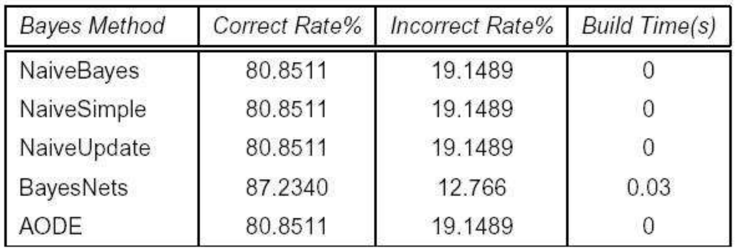
Bayes Method	Correct Rate%	Incorrect Rate%	Build Time(s)
NaiveBayes	80.8511	19.1489	0
NaiveSimple	80.8511	19.1489	0.02
NaiveUpdate	80.8511	19.1489	0
BayesNets	78.7234	21.2766	0
AODE	78.7234	21.2766	0.03

15 Attributes



Bayes Method	Correct Rate%	Incorrect Rate%	Build Time(s)
NaiveBayes	80.8511	19.1489	0
NaiveSimple	80.8511	19.1489	0
NaiveUpdate	80.8511	19.1489	0
BayesNets	80.8511	19.1489	0
AODE	78.7234	21.2766	0

11 Attributes



Bayes Method	Correct Rate%	Incorrect Rate%	Build Time(s)
NaiveBayes	80.8511	19.1489	0
NaiveSimple	80.8511	19.1489	0
NaiveUpdate	80.8511	19.1489	0
BayesNets	87.2340	12.766	0.03
AODE	80.8511	19.1489	0

Discussion

Some interesting points to be discussed.

- Overfitting? (AODE with 11 attributes - 88% correct with 82% training) BayesNet results listed table below.
- Discretize settings.
- Find a better attribute subset?

Training%	*Testing%*	*Correct Rate%*
66	34	87.2340
60	40	82.1429
75	25	85.7143
82	18	84.0000
90	10	78.5714

Ending

Questions regarding Analysis results?

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

E-mail: xiao@cs.queens.ca

Thank you