Discretization Methods for Classification

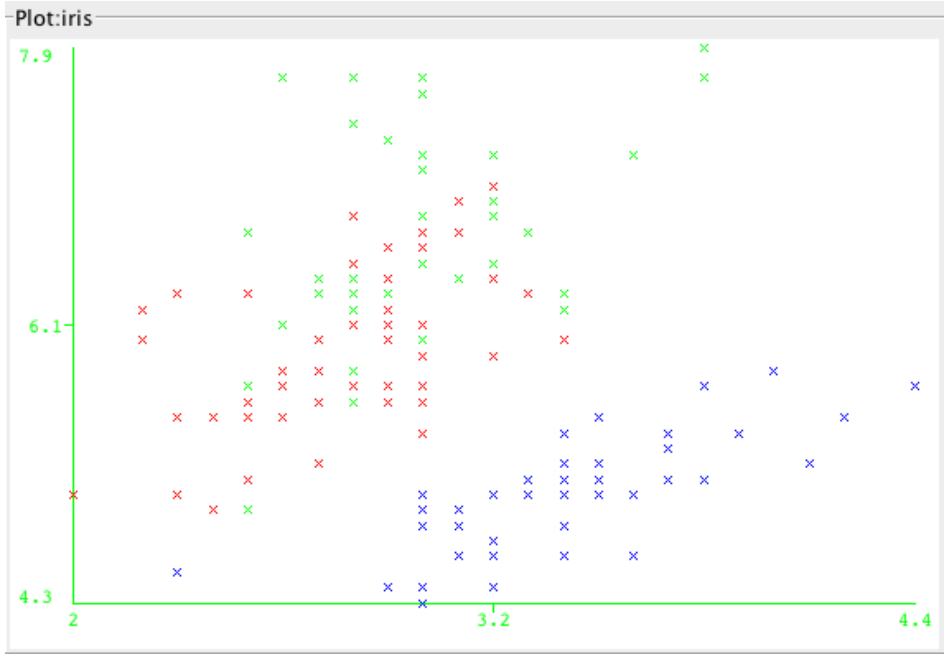
Alysander Stanley, Kevin B. Korb & Ann E. Nicholson
Clayton School of Information Technology
Monash University

{alysander@gmail.com,kbkorb@gmail.com,ann.nicholson@monash.edu}

Discretization Methods for Classification

Contents

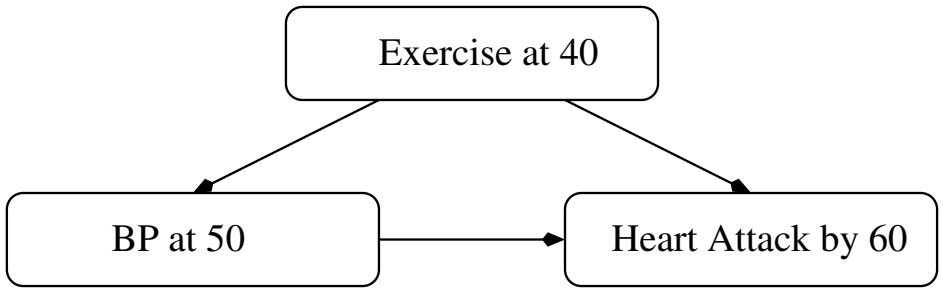
Classification



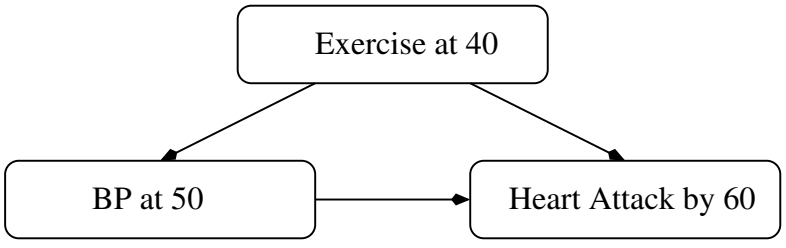
Iris: Sepal length (Y) v Sepal width (X)

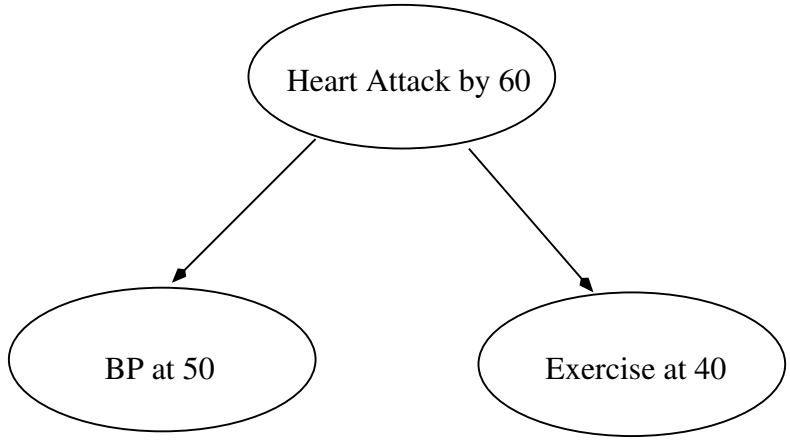
- Unsupervised: Find clusters in the instance space
- Supervised: Find predictors of known classes (colors above)

Causal Bayesian Networks



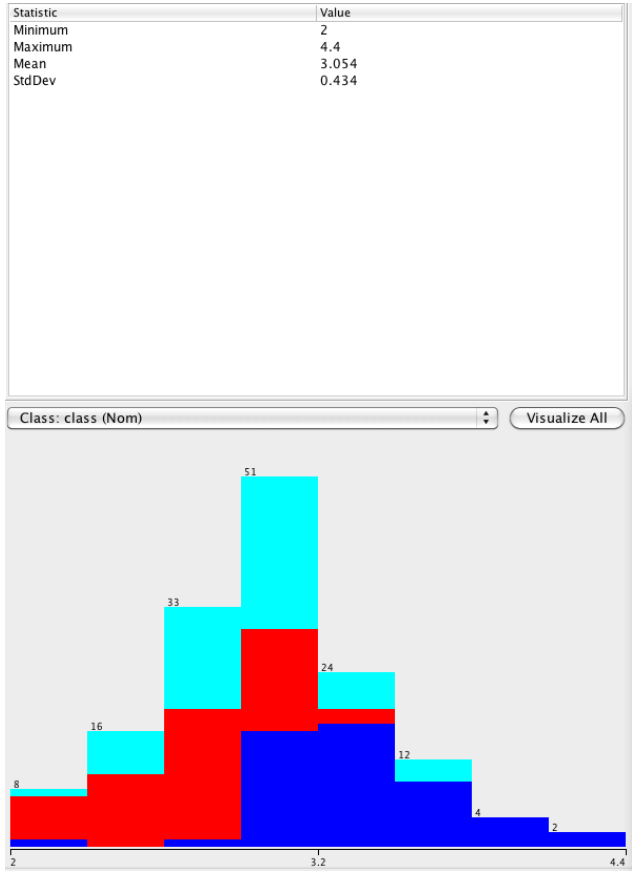
Naive Bayes Models





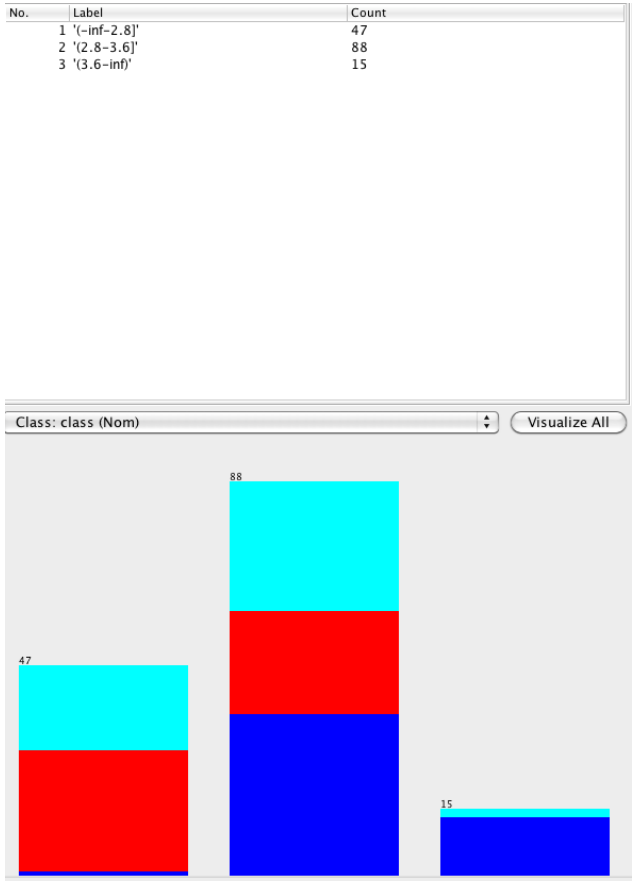
Discretization

The most basic case is taking a single variable (dimension) in isolation and finding a useful division into intervals. E.g., Iris Sepal Width



Discretization

The most basic case is taking a single variable (dimension) in isolation and finding a useful division into intervals. E.g., Iris Sepal Width



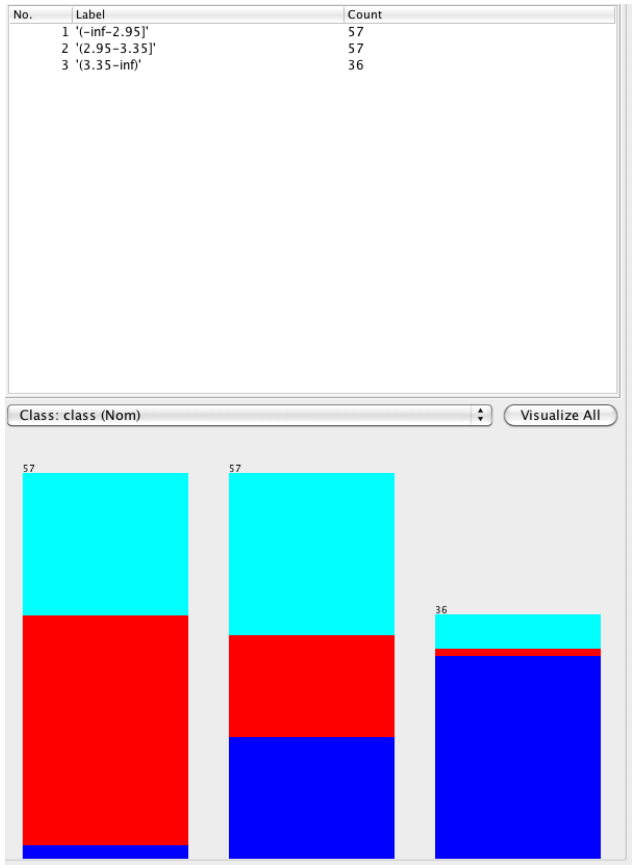
Discretization Techniques

Some techniques:

- Equal Frequency Binning (EFB3/5)
- Maximum Likelihood (Predictive Accuracy; PA Fit)
- Maximizing Area Under the Curve (AUC Fit)
- Enropy MDL (Ent-MDL)

Entropy-MDL Discretization

Uses MDL to optimize the number of intervals; maximizes entropy in the classification given those intervals. E.g., Iris Sepal Width



Evaluating Classifiers

Evaluation

Having found a discretization using a training set, evaluate the discretization using a (different) test set.

Predictive Accuracy

	T (p)	F (1-p)
“T”	TP	FP
“F”	FN	TN

Predictive Accuracy

	T (p)	F (1-p)
"T"	0.9	0.2
"F"	0.1	0.8

$$\begin{aligned} PA &= p(0.9) + (1-p)(0.8) \\ &= 1 - (p \times 0.1 + (1-p)0.2) \end{aligned}$$

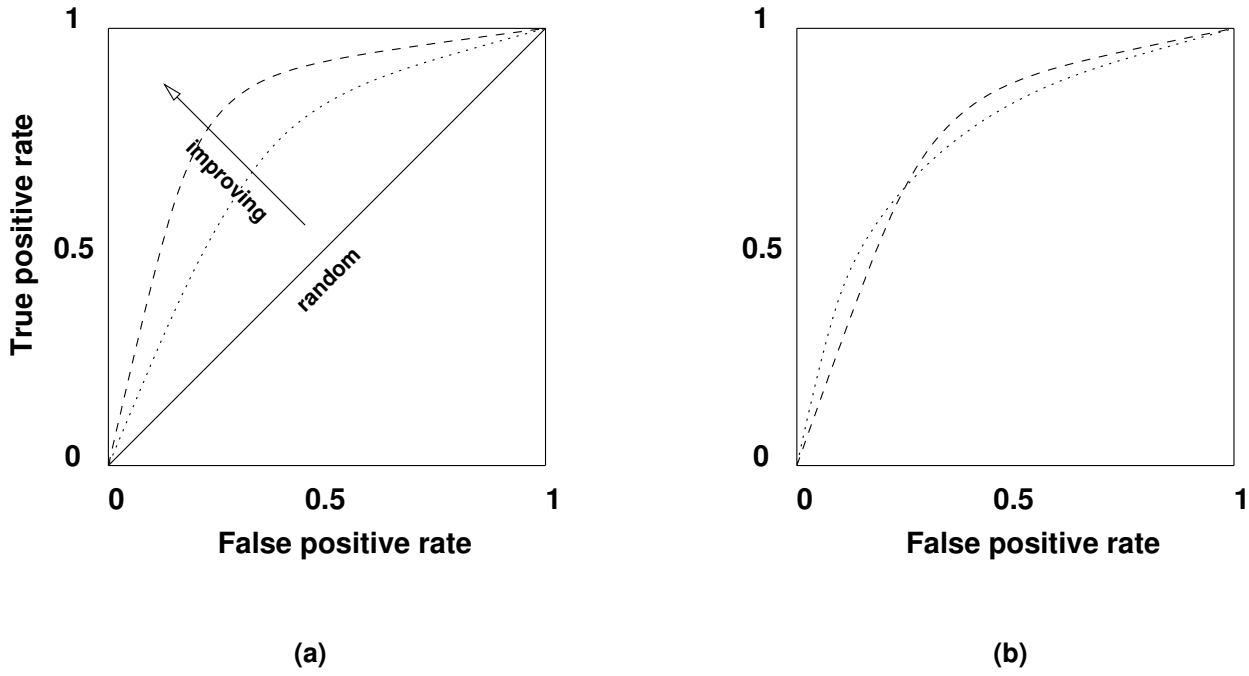
Predictive Accuracy

	Edible	Poison
“Edible”		y
“Poison”	x	

But, $v(x) \neq v(y)$

ROC

Maps TPRs to FPRs $P('T'|T)$ *v* $P('T'|F)$:



- PA (for FPR): $p \times TPR + (1 - p)(1 - FPR)$
- AUC: Integrate under the curve

Cost-Based Classification

Classify (discretize) to maximize expected value of classification. I.e.,

	T (p)	F (1-p)
"T"	tp, u(tp)	fp, u(fp)
"F"	fn, u(fn)	tn, u(tn)

$$\max \sum_i (tp \times u(tp) + fp \times u(fp) + fn \times u(fn) + tn \times u(tn))$$

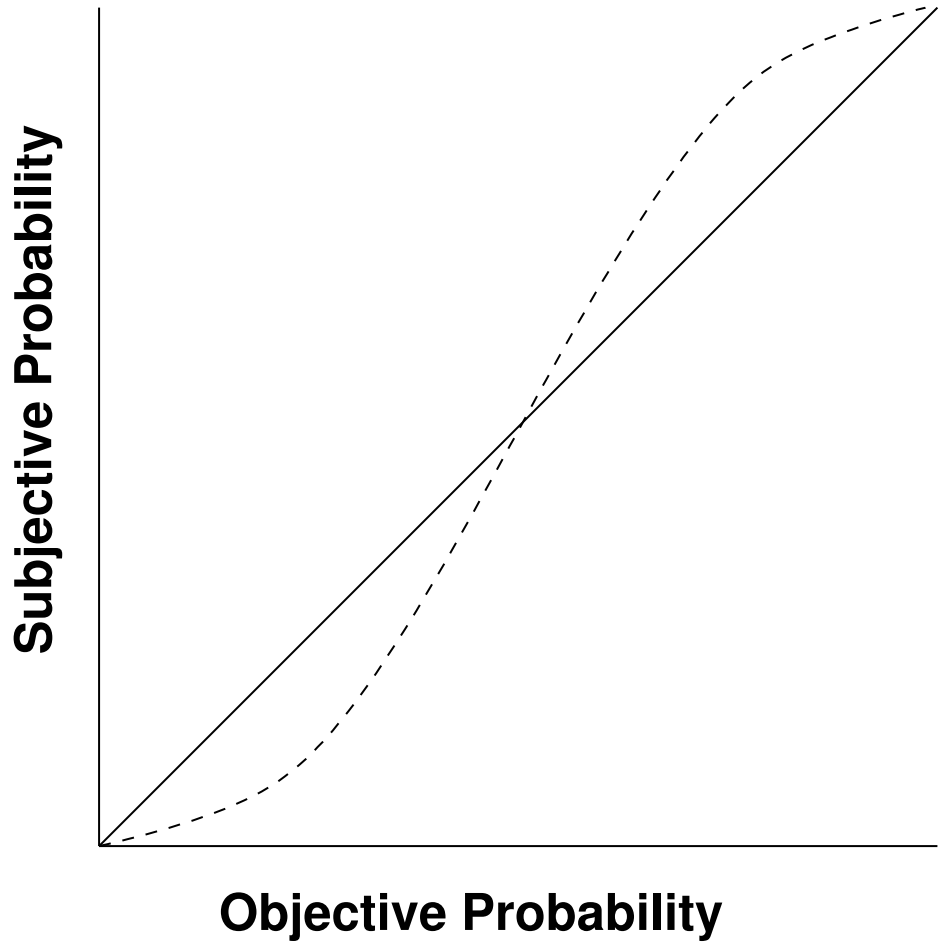
This *ought* to be the gold standard for classification, when turning from training to test sets!

Cost-Based Evaluation

I.e., the Bayesian gold standard for evaluation is/ought to be maximizing expected value in test sets:

$$\max \sum_i (tp \times u(tp) + fp \times u(fp) + fn \times u(fn) + tn \times u(tn))$$

Calibration



Calibration

Why does it matter?

Consider:

- P(“Edible”) = 0.99
- P(“Edible”) = 0.51

They both count the same on PA.

Clearly, cost-based evaluation takes this into account. We would like an evaluation measure that also does so when utilities aren’t available.

Bayesian Information Reward

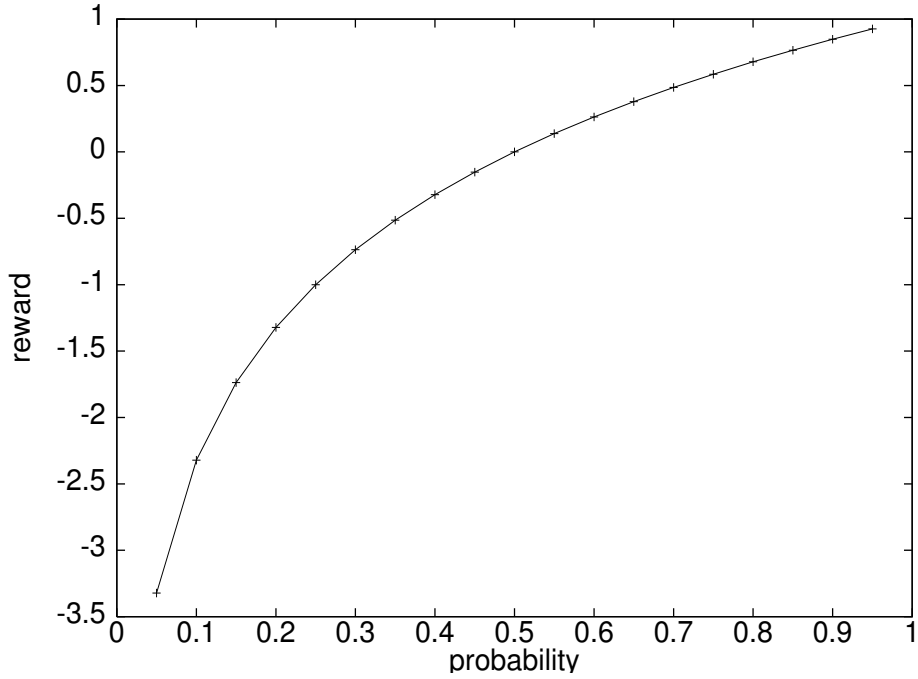
Bayesian information reward (Good, 1952; Korb & Nicholson, 2011) is a log scoring rule that provides a probability-weighted reward for every test instance which simultaneously

- Rewards classification accuracy
- Calibration

Good's Binomial Information Reward

$$IR_G = \sum_i [1 + \log_2 P(c_i)]$$

where c_i is the actual class.



Bayesian Information Reward

BIR generalizes this by:

- Generalizing to multinomial classes $\{C = c_i\}, P(c_i) = \hat{p}_i$
- Relativizing reward for $\hat{p}_i$ to the prior probability p_i

$$BIR = \frac{1}{n} \sum_{i,j} \log \frac{\hat{p}_i}{p_i} + \log \frac{1 - \hat{p}_j}{p_j}$$

where n is arity, i indexes true classes and j indexes false classes.

GA-Slicer

- GA search for the optimal multivariate discretization
  - Weka plugin
- Classifiers: J48, NB, AODE
  - Seeded with random discretizations (1-3 cutpoints)
- Reproduction:
  Crossover (0.25) XOR clone & mutate (0.75)
- Optimizing: PA, AUC, Cost, BIR
  - Supervised evaluation using cross validation

Results

Table: Mean BIR $\pm$ sd using AODE (**bold** = best result)

Dataset	BIR FIT	PA FIT	AUC FIT	MDL	EFB3
bupa	0.04 ± 0.03	0.06 ± 0.02	0.06 ± 0.03	0.04 ± 0.03	$\mathbf{0.08 \pm 0.02}$
credit-a	$\mathbf{0.33 \pm 0.03}$	0.32 ± 0.03	$\mathbf{0.33 \pm 0.03}$	0.31 ± 0.03	$\mathbf{0.33 \pm 0.03}$
credit-g	0.08 ± 0.02	$\mathbf{0.09 \pm 0.02}$	0.08 ± 0.02	0.08 ± 0.02	0.08 ± 0.02
heart-statl	$\mathbf{0.29 \pm 0.04}$	0.27 ± 0.05	$\mathbf{0.29 \pm 0.05}$	0.26 ± 0.06	$\mathbf{0.29 \pm 0.05}$
heart-c	$\mathbf{0.30 \pm 0.06}$	0.28 ± 0.07	0.28 ± 0.06	0.28 ± 0.06	0.28 ± 0.06
heart-h	0.24 ± 0.11	0.24 ± 0.11	0.23 ± 0.12	$\mathbf{0.25 \pm 0.12}$	0.24 ± 0.12
diabetes	$\mathbf{0.14 \pm 0.02}$	0.12 ± 0.03	$\mathbf{0.14 \pm 0.03}$	0.12 ± 0.03	0.13 ± 0.03
hepatitis	0.05 ± 0.10	0.09 ± 0.09	0.08 ± 0.11	0.08 ± 0.08	0.08 ± 0.09
labor-neg	0.48 ± 0.05	0.47 ± 0.08	0.43 ± 0.07	0.36 ± 0.11	$\mathbf{0.51 \pm 0.06}$
adult	0.21 ± 0.00	0.20 ± 0.01	0.21 ± 0.00	$\mathbf{0.22 \pm 0.00}$	0.18 ± 0.00
echocard	-0.01 ± 0.09	0.02 ± 0.06	0.00 ± 0.05	-0.01 ± 0.05	$\mathbf{0.07 \pm 0.07}$
ionosph	$\mathbf{0.30 \pm 0.10}$	0.25 ± 0.14	0.29 ± 0.12	0.09 ± 0.18	0.16 ± 0.15
sonar	0.08 ± 0.11	0.12 ± 0.18	$\mathbf{0.15 \pm 0.16}$	-0.11 ± 0.22	0.07 ± 0.18
non-binomials					
Avg.	$\mathbf{0.22 \pm 0.04}$	0.21 ± 0.05	0.20 ± 0.06	0.19 ± 0.05	0.21 ± 0.05

Results

Table: Significance tests for BIR v Others with AODE (**bold** = sig result)

	BIR	**PA**	**AUC**
BIR vs AUC	8 – 0 – 5 (0.88)	9 – 0 – 4 (0.40)	7 – 0 – 6 (0.83)
BIR vs PA	**19 – 0 – 5 (<0.05)**	13 – 0 – 11 (0.74)	9 – 0 – 4 (0.40)
BIR vs MDL	**16 – 0 – 8 (<0.05)**	11 – 0 – 13 (0.94)	**10 – 0 – 3 (<0.05)**
BIR vs EFB3	19 – 0 – 5 (0.061)	15 – 0 – 9 (0.43)	5 – 0 – 8 (0.16)

Results

Table: Mean BIR $\pm$ sd using NB (**bold** = best result)

Dataset	**BIR FIT**	**PA FIT**	**AUC FIT**	**MDL**	**EFB3**
bupa	0.04 ± 0.03	0.04 ± 0.03	0.03 ± 0.03	0.03 ± 0.03	$\mathbf{0.05 \pm 0.02}$
credit-a	$\mathbf{0.29 \pm 0.05}$	0.28 ± 0.05	0.28 ± 0.05	0.25 ± 0.04	0.25 ± 0.04
credit-g	0.06 ± 0.02	$\mathbf{0.07 \pm 0.03}$	$\mathbf{0.07 \pm 0.02}$	$\mathbf{0.07 \pm 0.03}$	0.06 ± 0.02
heart-statl	$\mathbf{0.29 \pm 0.04}$	0.25 ± 0.04	0.27 ± 0.07	0.21 ± 0.08	0.24 ± 0.07
heart-c	$\mathbf{0.28 \pm 0.07}$	0.24 ± 0.08	0.26 ± 0.07	0.22 ± 0.08	0.21 ± 0.08
heart-h	$\mathbf{0.24 \pm 0.10}$	0.23 ± 0.11	0.22 ± 0.13	0.21 ± 0.14	0.18 ± 0.16
diabetes	$\mathbf{0.15 \pm 0.03}$	0.12 ± 0.03	0.13 ± 0.03	0.10 ± 0.04	0.10 ± 0.04
hepatitis	-0.04 ± 0.10	-0.03 ± 0.13	-0.04 ± 0.11	$\mathbf{-0.01 \pm 0.09}$	-0.08 ± 0.10
labor-neg	0.49 ± 0.07	0.48 ± 0.07	0.41 ± 0.09	0.37 ± 0.11	$\mathbf{0.52 \pm 0.07}$
adult	$\mathbf{0.15 \pm 0.01}$	0.14 ± 0.01	0.13 ± 0.01	$\mathbf{0.15 \pm 0.01}$	0.12 ± 0.01
echocard	-0.03 ± 0.09	0.00 ± 0.10	0.00 ± 0.08	-0.04 ± 0.09	$\mathbf{0.07 \pm 0.09}$
ionosph	$\mathbf{0.29 \pm 0.09}$	0.27 ± 0.11	0.27 ± 0.09	$-\infty \pm \mathit{NaN}$	-0.19 ± 0.35
sonar	$\mathbf{0.01 \pm 0.18}$	-0.17 ± 0.18	-0.11 ± 0.28	-0.26 ± 0.28	-0.36 ± 0.29
non-binomials					
Avg.	$\mathbf{0.19 \pm 0.05}$	0.16 ± 0.06	0.15 ± 0.08	$-\infty \pm \mathit{NaN}$	0.08 ± 0.07

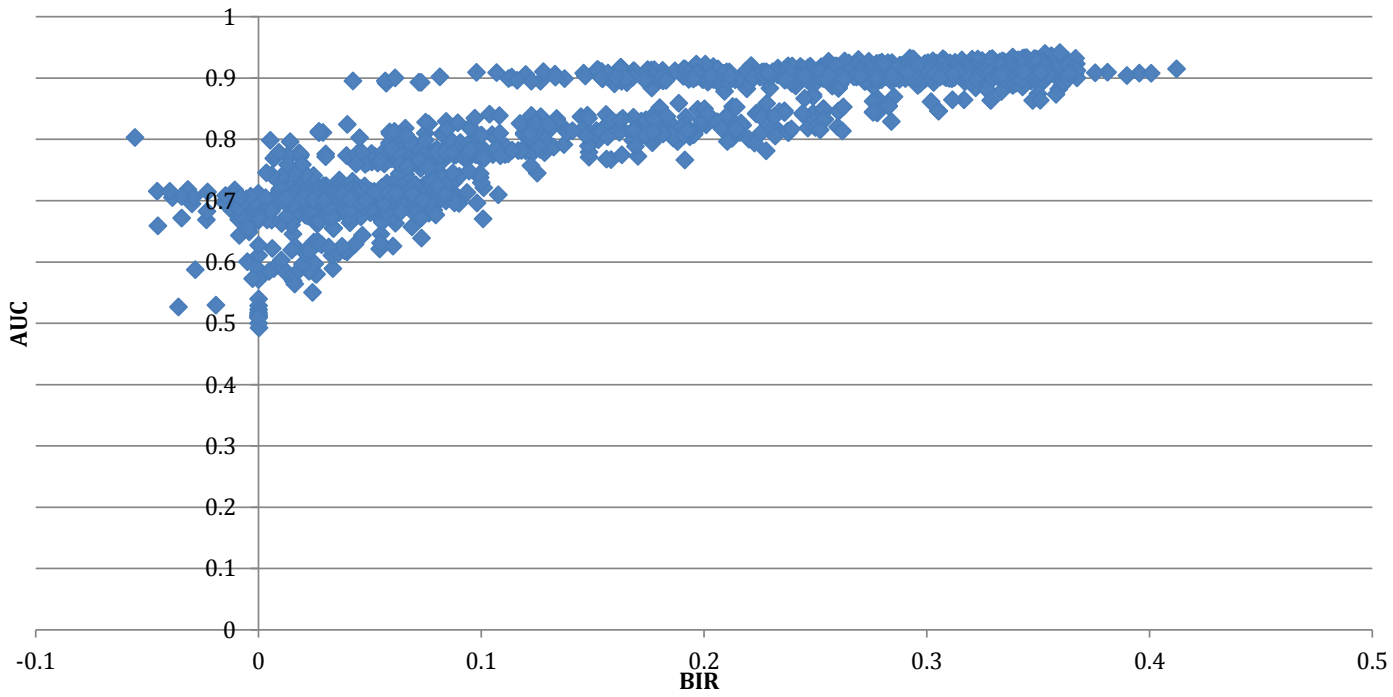
Results

Table: Significance tests for BIR v Others with NB (**bold** = sig result)

	BIR	**PA**	**AUC**
BIR vs AUC	**11 – 0 – 2 (<0.05)**	5 – 0 – 8 (0.88)	7 – 0 – 6 (0.36)
BIR vs PA	**21 – 0 – 3 (<0.05)**	12 – 0 – 12 (0.60)	8 – 0 – 5 (0.12)
BIR vs MDL	**17 – 2 – 5 (<0.05)**	12 – 0 – 12 (0.50)	8 – 0 – 5 (0.12)
BIR vs EFB3	**18 – 0 – 6 (<0.05)**	15 – 0 – 9 (0.06)	5 – 0 – 8 (0.62)

BIR v AUC

Pearson's $r = 0.8825$



Future Work

- Complete the analysis! (Esp cost-based discretization)
- Publish
- Donate the result to Weka

References

J Pearl (1988). *Probabilistic Reasoning in Intelligent Systems.* San Mateo, CA: Morgan Kaufmann.

C Twardy, A Nicholson, K Korb and J McNeil (2006). Epidemiological data mining of cardiovascular Bayesian networks. *Electronic Jrn of Health Informatics, 1,* 1-13.

K Irani and U Fayyad (1993). Multi-interval discretization of continuous valued attributes for classification learning. In *Proc of the 13th Joint Conf on AI,* 1022-1027.

S Bay (2000). Multivariate discretization of continuous variables for set mining. In *Proc of Knowledge Discovery and Data Mining.*

P Turney (1995). Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of AI Research,* 369-409.

K Korb, L Hope, and M Hughes (2001). The evaluation of predictive learners: Some theoretical and empirical results. *European Conference on Machine Learning,* 276-287.

K Korb and A Nicholson (2010). *Bayesian Artificial Intelligence,* 2nd ed. CRC/Chapman Hall.