

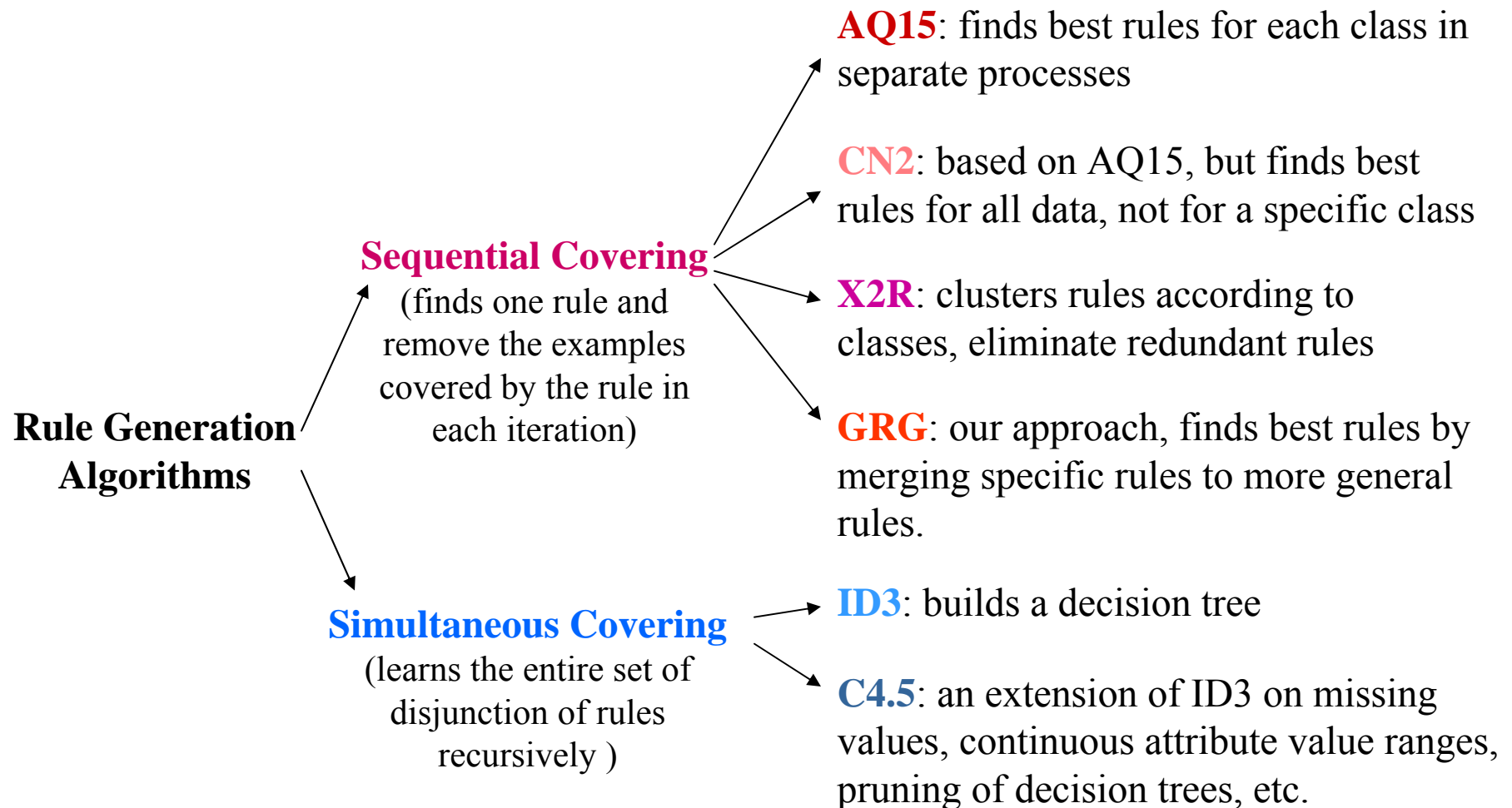
# Project Information Page

- **Project type:** Honours Year Project
- **Project Area:** Programming Languages, Algorithms, Theory
- **Project Title:** Rule Generation from Discrete Data
- **Project No.:** H001090
- **Student Name:** Gong Tianxia
- **Project Advisor:** Prof. Rudy Setiono
- **Date of Completion:** April, 2006
- **Deliverables:** Report 1 Volume, Program 1CD
- **Software:** Eclipse, WEKA, MS Windows XP
- **Hardware:** Intel Pentium 3.0GHz CPU, Kingston 512 RAM

# Introduction

- Rule set generation is one of the most expressive and human readable methods in machine learning
- Our aim is to find a rule generation method that have:
  - High accuracy
  - Low time complexity
  - Small rule list size
- We modified and implemented Greedy Rule Generation algorithm, re-implemented Chi2 discretization, entropy and information gain measures.
- We conducted intensive testing on real world data and compared the result to other machine learning approaches.

# Related Works



# Greedy Rule Generation (GRG)

GRG(*Examples*, *Attributes*)

*optimal\_rule\_list*  $\leftarrow$  {}

*subspaces*  $\leftarrow$  Build\_subspaces(*Examples*, *Attributes*)

*subspaces*  $\leftarrow$  Label\_subspaces(*Examples*, *Attributes*)

while (*subspace* *s* such that Label(*s*)  $\neq$  “unlabeled”)

*rule\_list*  $\leftarrow$  {}

$\bar{R}$   $\leftarrow$  the rule covers most examples in *subspaces*

*rule\_list*  $\leftarrow$  + {other rules such that the class of is the same as  $\bar{R}$  or “unlabeled”}

while (Mergeable(*rule\_list*))

*rule\_list*  $\leftarrow$  *rule\_list* + Merge\_rule(*rule\_list*)

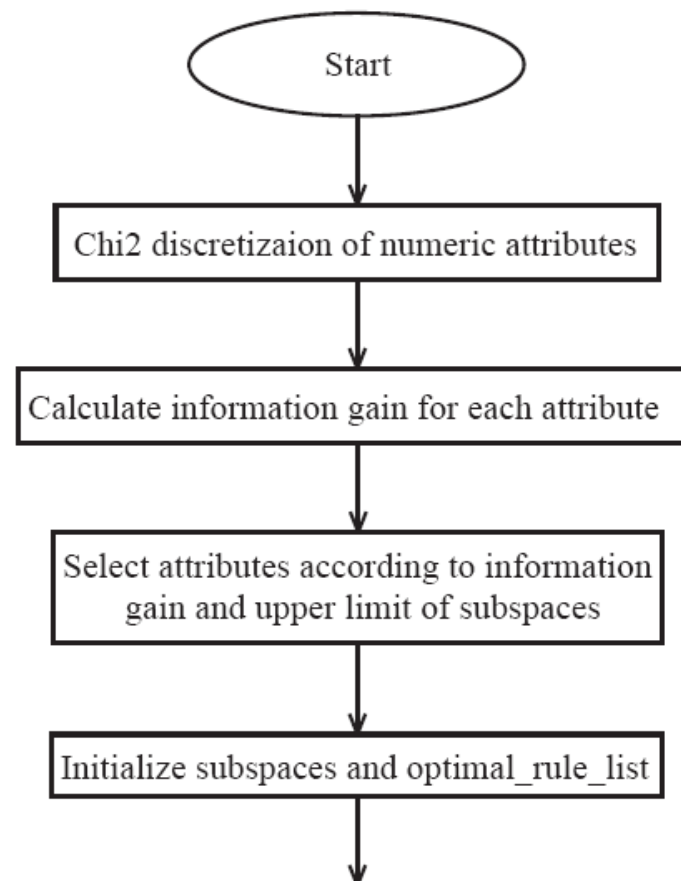
$R^*$   $\leftarrow$  Find\_best\_rule(*rule\_list*)

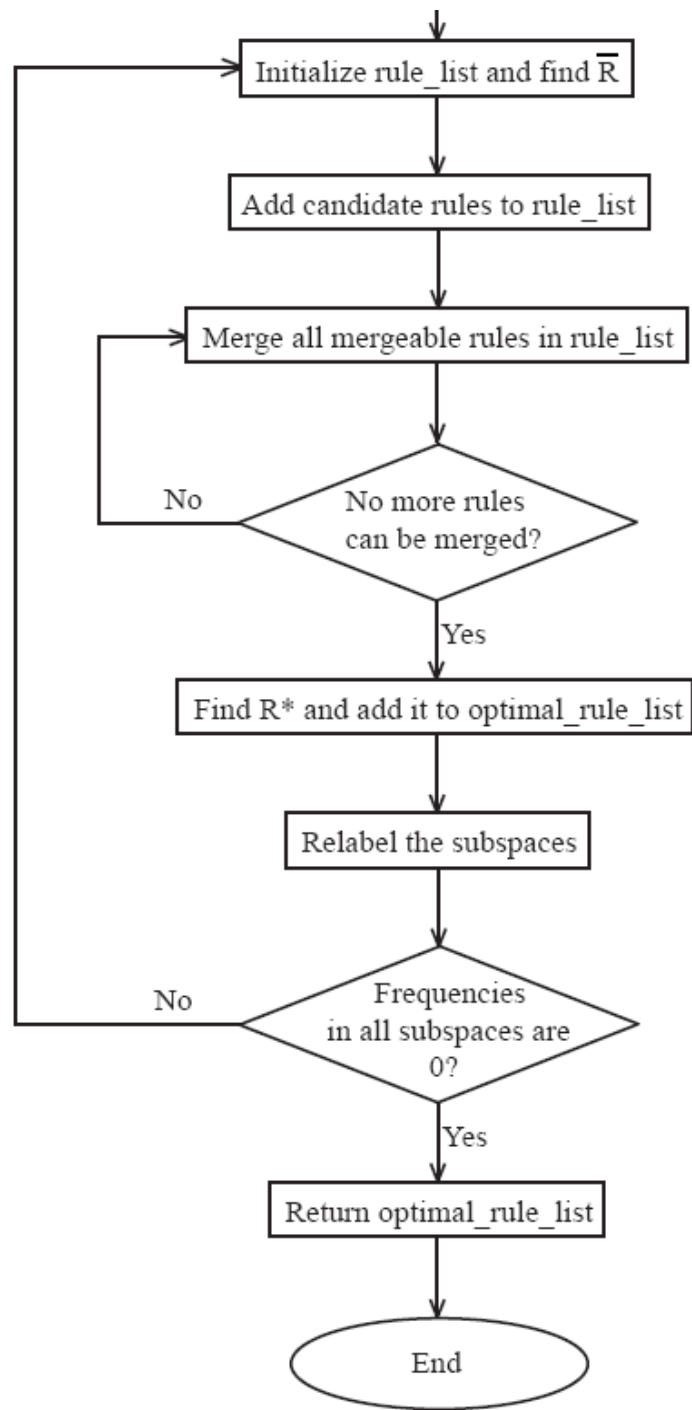
*subspaces*  $\leftarrow$  Relabel\_subspaces(*Examples*, *Attributes*,  $R^*$ )

*optimal\_rule\_list*  $\leftarrow$  *optimal\_rule\_list* +  $R^*$

return *optimal\_rule\_list*

# Program Flow





# Experiment Setup

- Ten 10-fold cross-validation scheme
- Chi2 discretization (Inconsistency rate = 5%)
- Subspace upper limit is set to 300
- heuristic of attribute selection: Entropy and Information Gain
- Missing Values:
  - A missing numeric value is replaced by average
  - A missing discrete value is replaced by a new value “unknown”

# Experiment Results

Dataset	10 ten-fold accuracy (%)	Number of rules	Number of useful attributes
australian	$85.68 \pm 0.31$	$5.69 \pm 0.40$	$3.00 \pm 0.00$
breast-cancer	$73.52 \pm 0.60$	$5.51 \pm 0.21$	$2.00 \pm 0.00$
breast-w	$95.43 \pm 0.32$	$6.30 \pm 0.85$	$3.00 \pm 0.00$
german	$72.47 \pm 0.62$	$3.64 \pm 0.33$	$2.00 \pm 0.00$
glass (G2)	$82.82 \pm 1.78$	$7.53 \pm 0.62$	$3.00 \pm 0.00$
heart-c	$82.44 \pm 0.94$	$9.74 \pm 0.37$	$3.98 \pm 0.04$
heart-h	$79.03 \pm 0.97$	$4.78 \pm 1.03$	$2.11 \pm 0.10$
heart-statlog	$81.15 \pm 0.77$	$14.40 \pm 0.18$	$4.72 \pm 0.13$
hepatitis	$79.45 \pm 2.14$	$3.37 \pm 0.20$	$4.11 \pm 0.10$
horse-colic	$82.10 \pm 0.35$	$3.61 \pm 0.21$	$2.00 \pm 0.00$
iris	$94.80 \pm 0.82$	$3.99 \pm 0.37$	$1.51 \pm 0.09$
lymphography	$80.38 \pm 1.72$	$15.50 \pm 1.03$	$5.00 \pm 0.00$
primary-tumor	$40.92 \pm 0.86$	$23.87 \pm 0.32$	$5.00 \pm 0.00$
sick	$97.61 \pm 0.04$	$4.10 \pm 0.11$	$6.00 \pm 0.00$

# Comparison to Decision Tree and Neural Network Approaches

Dataset	GRG	C5.0	M5'	N2C2S
australian	85.68 ± 0.31	85.30 ± 0.50	85.80 ± 0.90	84.80 ± 0.70
breast-cancer	73.52 ± 0.60	73.30 ± 1.60	69.60 ± 2.30 •	67.80 ± 2.10 •
breast-w	95.43 ± 0.32	94.50 ± 0.30 •	95.30 ± 0.30	96.50 ± 0.20 ◊
german	72.47 ± 0.62	71.20 ± 1.00	72.90 ± 0.70	70.10 ± 1.60 •
glass (G2)	82.82 ± 1.78	78.70 ± 2.10 •	81.80 ± 2.20	77.90 ± 2.50 •
heart-c	82.44 ± 0.94	76.80 ± 1.40 •	80.90 ± 1.40	82.40 ± 1.90
heart-h	79.03 ± 0.97	79.80 ± 0.90	79.00 ± 0.80	81.60 ± 1.60 ◊
heart-statlog	81.15 ± 0.77	78.70 ± 1.40 •	82.20 ± 1.00	77.50 ± 1.00 •
hepatitis	79.45 ± 2.14	79.30 ± 1.20	81.90 ± 2.20	81.90 ± 3.30
horse-colic	82.10 ± 0.35	85.30 ± 0.60 ◊	84.60 ± 0.70 ◊	78.90 ± 1.20 •
iris	94.80 ± 0.82	94.50 ± 0.70	94.70 ± 0.70	96.60 ± 1.60 ◊
lymphography	80.38 ± 1.72	75.40 ± 2.80 •	79.80 ± 1.40	82.90 ± 1.90 ◊
primary-tumor	40.92 ± 0.86	41.80 ± 1.30	45.10 ± 1.60 ◊	45.90 ± 1.30 ◊
sick	97.61 ± 0.04	98.80 ± 0.10 ◊	98.30 ± 0.10 ◊	97.50 ± 0.10

# Comparison to Other Rule Generation Approaches

Accuracy (%)	Lymphography	Breast Cancer	Primary Tumor
GRG	80	74	41
CN2*	78-82	70-71	36-37
AQ15 <sup>†</sup>	80-82	66-68	29-41
Human Experts	85 <sup>‡</sup>	64	42

	Lymphography		Breast Cancer		Primary Tumor	
	Selectors <sup>§</sup>	Rules	Selectors	Rules	Selectors	Rules
GRG	35	19	11	22	75	24
CN2**	-	12-24	-	4-28	-	19-42
AQ15	10-37	76	7-160	208	112-551	562

\* The accuracies and complexities varies because of the use of three different threshold in CN2: 90%, 95%, and 99%

<sup>†</sup> The accuracies and complexities varies because of the use of three different strategies in AQ15: no-truncation, eliminate-unique-cover and best-complex

<sup>‡</sup> Estimated ()

<sup>§</sup> Selector is defined in Section 2.2.1

\*\* The number of attributes (selectors) is not stated in the original paper

# Conclusion

- GRG has shown competitive performance in comparison with decision tree approaches C5.0 and M5', and neural network approach N2C2S.
- GRG has shown performance over rule generation algorithms as AQ and CN2 in terms of:
  - Higher accuracy
  - Lower time complexity
  - Smaller rule list size
- Future work for GRG improvement may include:
  - Optimization rule merging process