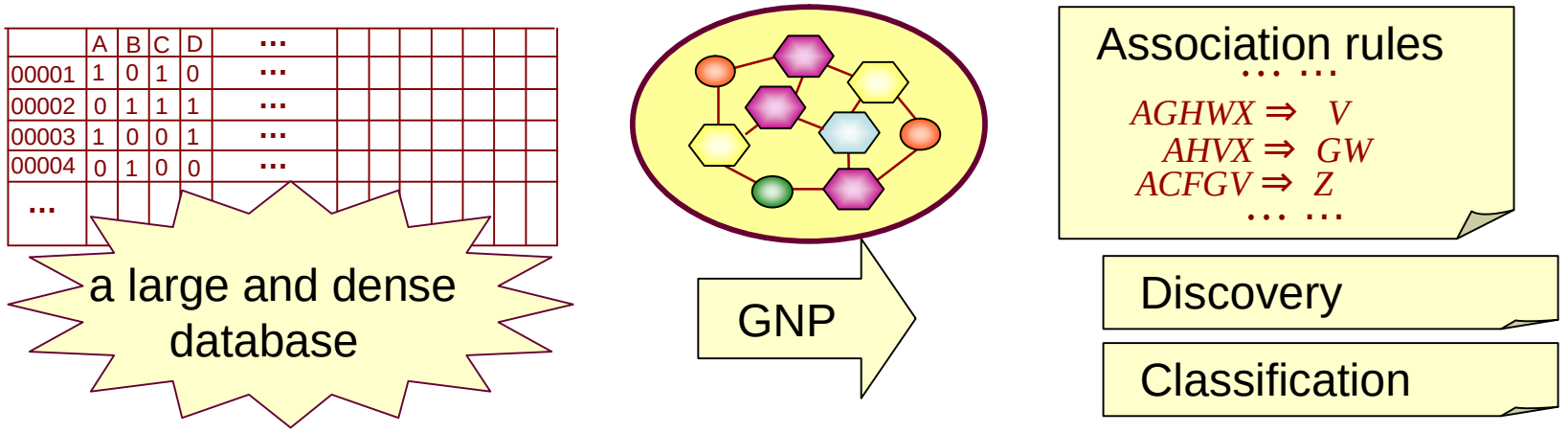
Application of Genetic Network Programming to Data Mining

Hirasawa Lab.

Objectives

1. Construct a system for data mining
   using Genetic Network Programming (GNP)

2. Study the properties of GNP
   when applying it to data mining



Contents

Association Rules

- Discovery of association relationships
  or correlations among a set of attributes
  in a database

- ***X* ⇒ *Y***
  Database tuples satisfying *X*
  are likely to satisfy *Y*
  X : antecedent, *Y* : consequent

- **support, confidence**
  support(*X*) = a/N
  support(*X*⇒*Y*) = b/N
  confidence(*X*⇒*Y*) = b/a

	Y	¬*Y*	Σ
X	b		a
¬*X*			
Σ			N

TID	*A*	*B*	*C*	*D*
1	1	0	1	0
2	0	1	1	1
3	1	1	1	1
4	0	1	0	1

$(A=1) \wedge (C=1) \Rightarrow (D=1)$

support = 0.25
confidence = 0.5

	D	¬*D*	Σ
AC	1		2
¬(*AC*)			
Σ			4

Support-confidence framework

- Finds all the rules meeting user-specified constrain such as minimum support or minimum confidence

TID	*A*	*B*	*C*	*D*
1	1	0	1	0
2	0	1	1	1
3	1	1	1	1
4	0	1	0	1
5	1	1	1	?

$(A=1) \wedge (C=1) \Rightarrow (D=1)$

support = 0.25

confidence = 0.5

TID	*A*	*B*	*C*	*D*
1	1	0	1	0
2	0	1	1	1
3	1	1	1	1
4	0	1	0	1
5	1	1	1	?

$(B=1) \wedge (C=1) \Rightarrow (D=1)$

support = 0.5

confidence = 1

Association Rule Mining

- (Attribute set X) ⇒ (Attribute set Y)

  Association relationships or
  correlations among a set of attributes

  Analysis, Discovery

$$(A_j = 1) \wedge \ldots \wedge (A_k = 1) \Rightarrow (A_m = 1) \wedge \ldots \wedge (A_n = 1)$$

TID	A_1	A_2	..			A_{n-1}	A_n
1	1	0				1	0
2	0	1				0	1
3	1	1				0	1
…						…	
N	0	1				0	1

- (Condition) ⇒ (Prediction)

  IF $(cond_1 \wedge cond_2 \wedge \ldots \wedge cond_4)$ *then (pred.)*

  Prediction, Classification

$$(A_j = 1) \wedge \ldots \wedge (A_k = 1) \Rightarrow (Z = 1)$$

  (Class association rule)
  (Fixed consequent)

TID	A_1	A_2	..			A_n	Z
1	1	0				0	0
2	0	1				1	1
3	1	1				1	0
…						…	
N	0	1				1	1

Association Rule Mining

Step 1. Generate rule candidates

(Attribute set X) ⇒ (Attribute set Y)

$$(A_j=1) \wedge \ldots \wedge (A_k=1) \Rightarrow (A_m=1) \wedge \ldots \wedge (A_n=1)$$

TID	A_1	A_2	..			A_{n-1}	A_n
1	1	0				1	0
2	0	1				0	1
3	1	1				0	1
…						…	
N	0	1				0	1

Step 2. Examine database

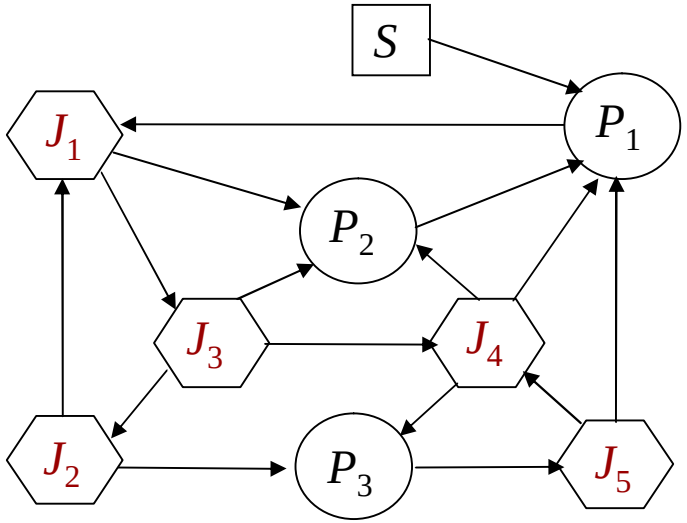
Calculate *support, confidence, …*

Step 3. Evaluation of rules

large computational overheads when the database is large and dense.

Genetic Network Programming (GNP)

- A kind of evolutionary methods
- Evolves arbitrary directed graph programs
- Includes
  Judgement nodes and
  Processing nodes
- Forms the optimal structure effectively



: Judgement node
(if-then type decision making function)

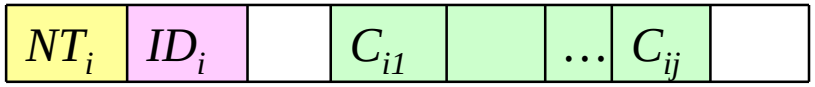
: Processing node
(action/processing function)

Basic Structure of GNP

Shingo Mabu, Kotaro Hirasawa and Jinglu Hu, "A Graph-Based Evolutionary Algorithm: Genetic Network Programming (GNP) and Its Extension Using Reinforcement Learning", *Evolutionary Computation*, MIT Press, Vol. 15, No. 3, pp. 369-398, 2007.

Gene structure of GNP

Gene structure of GNP (node i)



NT_i : node type

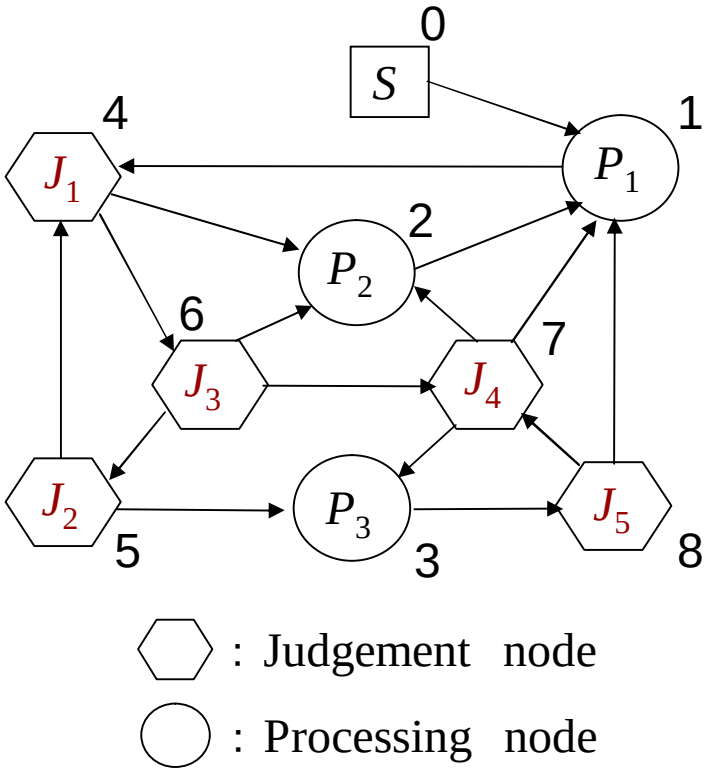
0: Start node 1: Processing node

2: Judgement node

ID_i : identification number

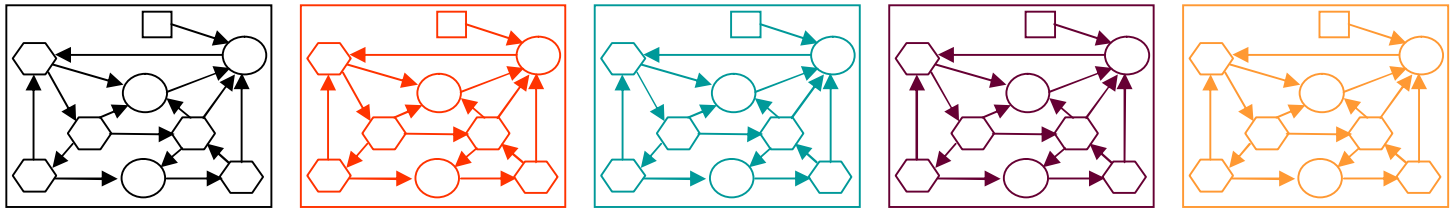
C_{ij} : node connection

0	0	0	1				
1	1	1	4				
2	1	2	1				
3	1	3	8				
4	2	1	2	6			
5	2	2	3	4			
6	2	3	2	5	7		
7	2	4	2	1	3		
8	2	5	1	7			



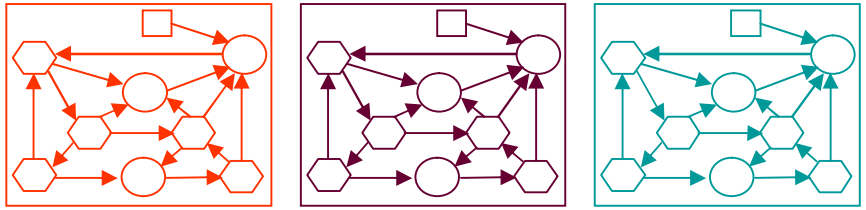
Selection and Reproduction

k^{th} generation



Ranked by Fitness value

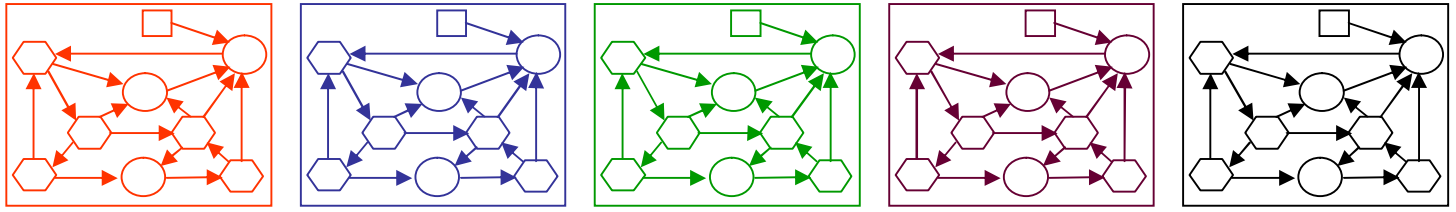
Selection



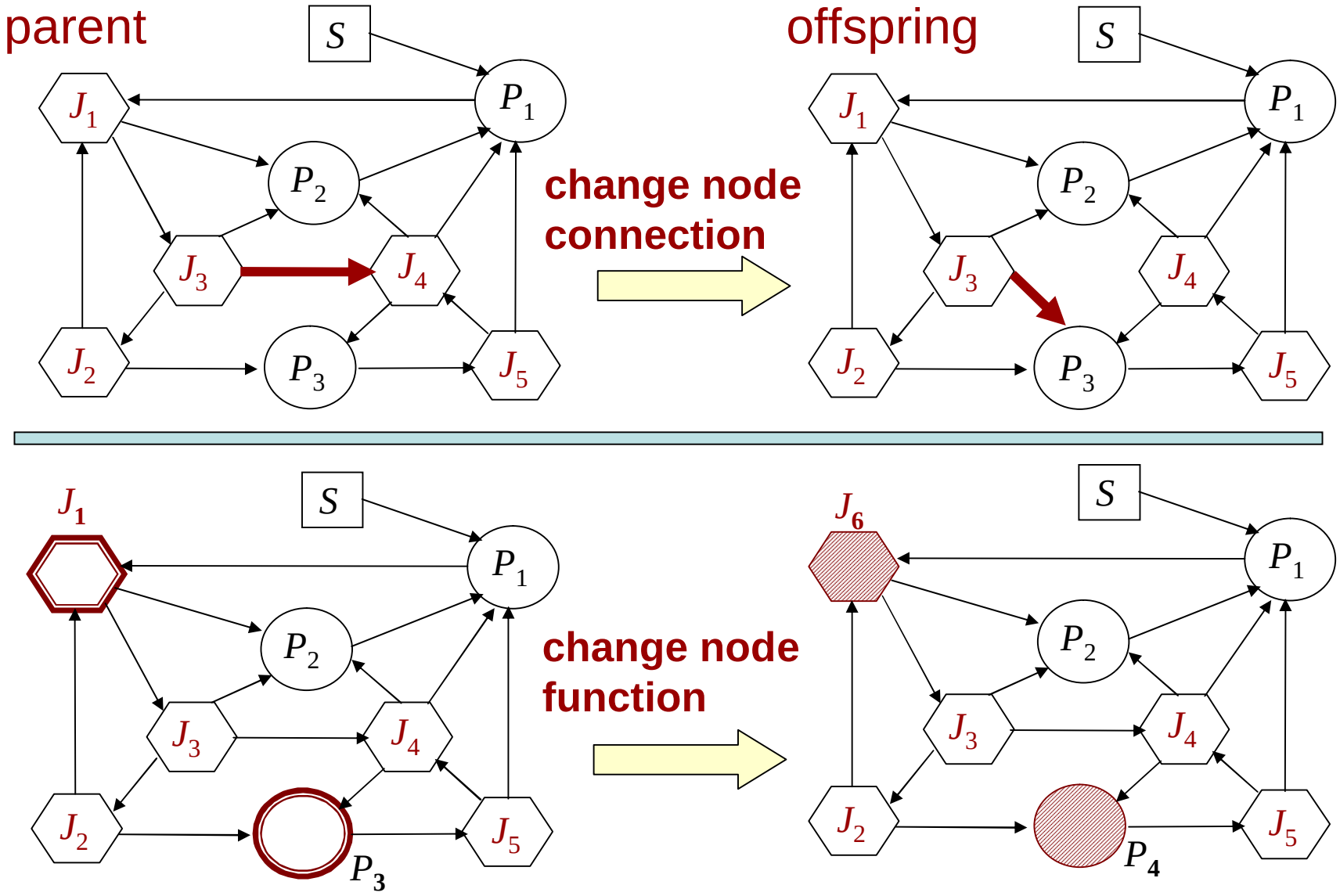
Genetic operations (mutation, crossover, elite)

Reproduction

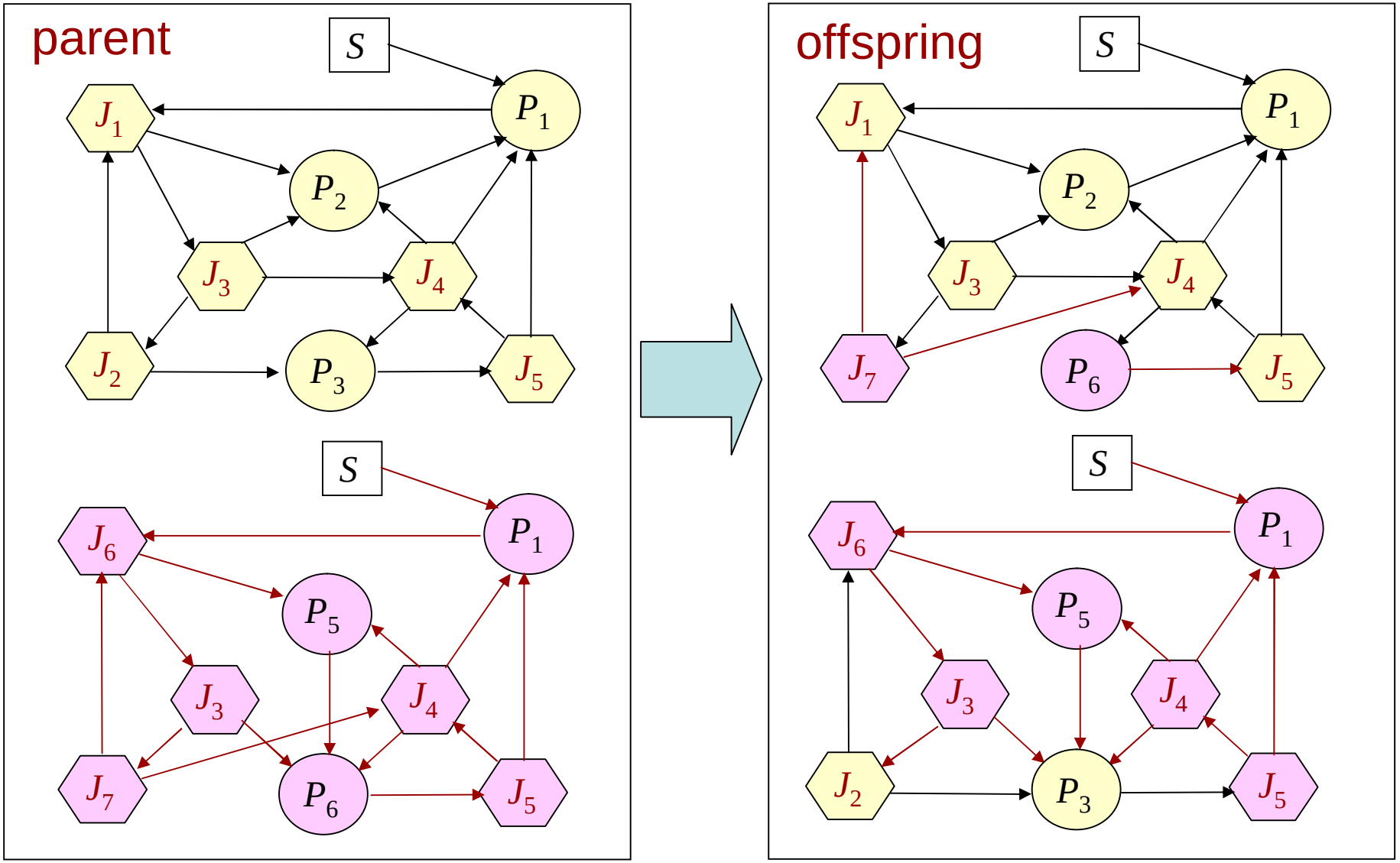
$k+1^{th}$ generation



Genetic Operation (mutation)

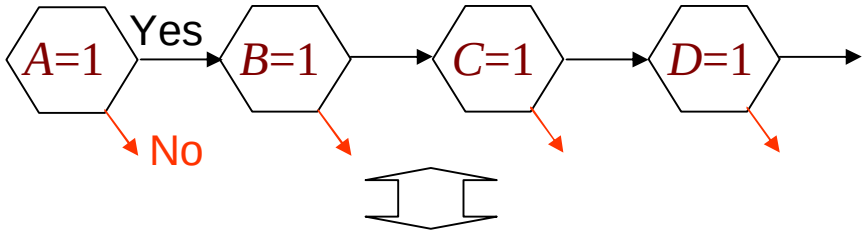


Genetic Operation (crossover)



The Basic Ideas

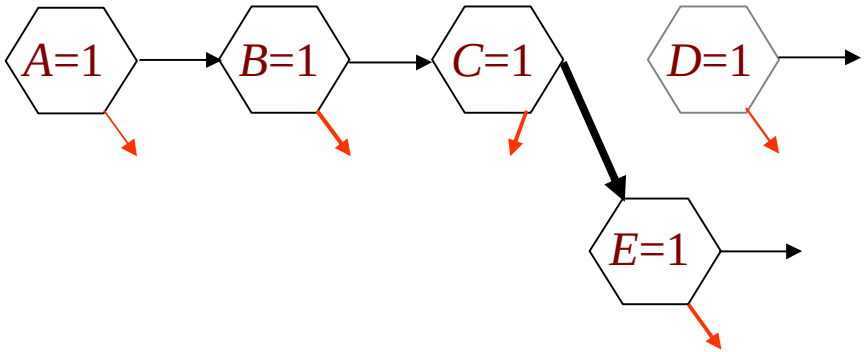
- Connect judgement nodes as association rules



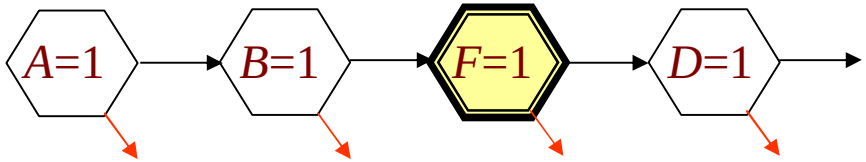
$$(A=1) \wedge (B=1) \wedge (C=1) \Rightarrow (D=1)$$
$$(A=1) \wedge (B=1) \Rightarrow (C=1) \wedge (D=1)$$
$$(A=1) \Rightarrow (B=1) \wedge (C=1) \wedge (D=1)$$

Represent connections of judgement nodes as association rules.

- Obtain candidates by genetic operations



$$(A=1) \wedge (B=1) \wedge (C=1) \Rightarrow (E=1)$$

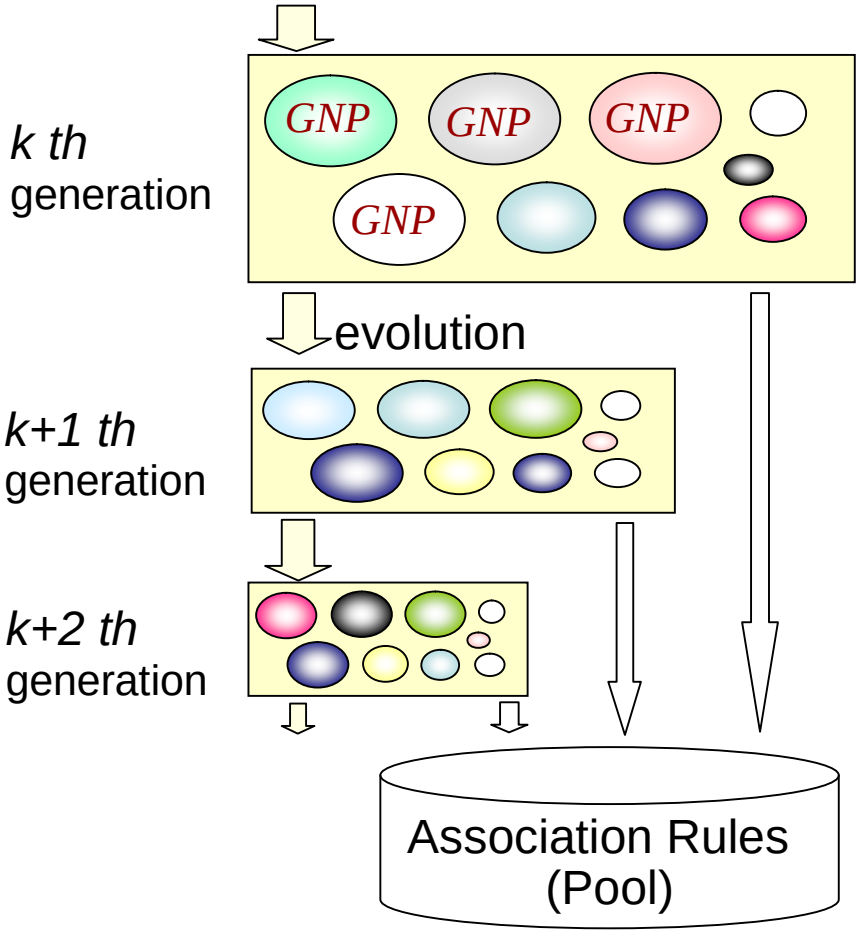


change the function of judgement node

$$(A=1) \wedge (B=1) \wedge (F=1) \Rightarrow (D=1)$$

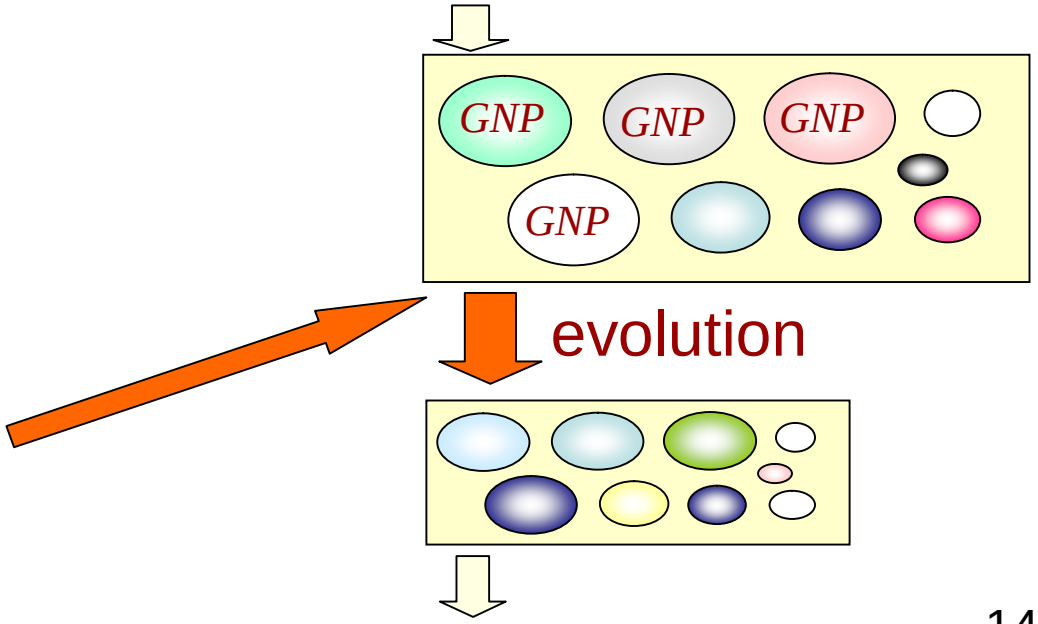
The Basic Ideas

☐ Extract rules through generations



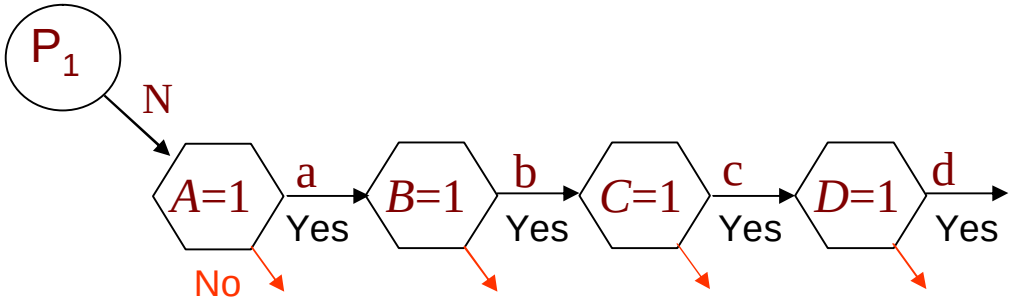
Evolve to store new rules in a pool, not to obtain an individual high fitness value

☐ Use of acquired information



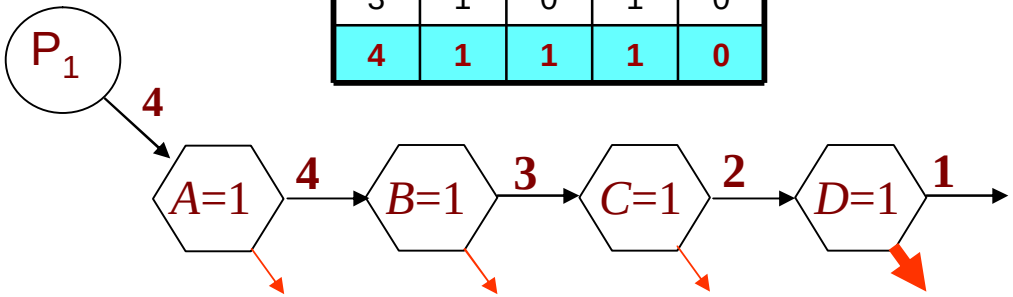
Extraction of Association Rules

- support, confidence



association rule	sup.	conf.
$A \Rightarrow B$	b / N	b / a
$A \Rightarrow B \wedge C$	c / N	c / a
$A \Rightarrow B \wedge C \wedge D$	d / N	d / a
$A \wedge B \Rightarrow C$	c / N	c / b
$A \wedge B \Rightarrow C \wedge D$	d / N	d / b
$A \wedge B \wedge C \Rightarrow D$	d / N	d / c

	A	B	C	D
1	1	1	1	1
2	1	1	0	1
3	1	0	1	0
4	**1**	**1**	**1**	**0**



association rule	sup.	conf.
$A \Rightarrow B$	3 / 4	3 / 4
$A \Rightarrow B \wedge C$	2 / 4	2 / 4
$A \Rightarrow B \wedge C \wedge D$	1 / 4	1 / 4
$A \wedge B \Rightarrow C$	2 / 4	2 / 3
$A \wedge B \Rightarrow C \wedge D$	1 / 4	1 / 3
$A \wedge B \wedge C \Rightarrow D$	1 / 4	1 / 2

chi-squared value of $X \Rightarrow Y$

☐ significance of the association

$$T = \sum_{AllCells} \frac{(Observatio\ n - Expectatio\ n)^2}{Expectatio\ n}$$

$$T = \frac{N(z - xy)^2}{xy(1-x)(1-y)}$$

significance level 5 % $T > 3.84$

significance level 1 % $T > 6.63$

Expectation (upper), Observation (lower)

	Y	$\neg Y$	Σ
X	Nxy Nz	$N(x\text{-}xy)$ $N(x\text{-}z)$	Nx
$\neg X$	$N(y\text{-}xy)$ $N(y\text{-}z)$	$N(1\text{-}x\text{-}y+xy)$ $N(1\text{-}x\text{-}y+z)$	$N(1\text{-}x)$
Σ	Ny	$N(1\text{-}y)$	N

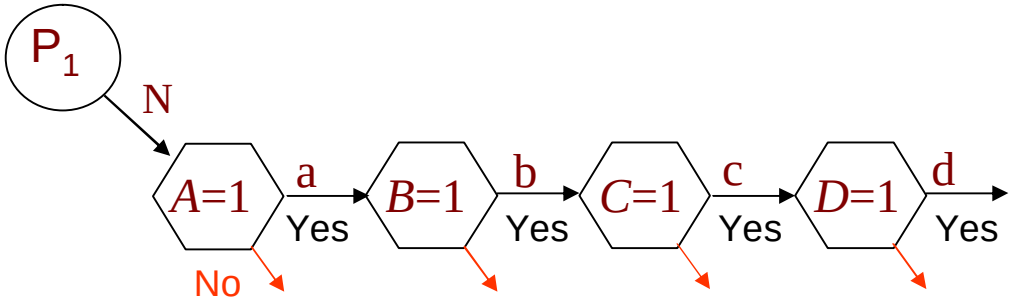
$support(X)=x$

$support(Y)=y$

$support(X \wedge Y)=z$

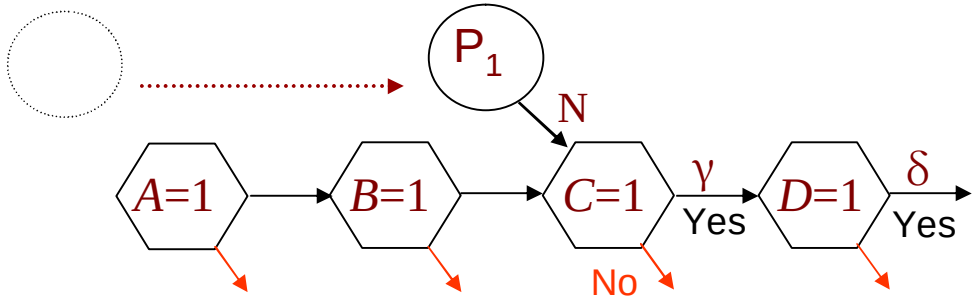
Extraction of Association Rules

- support, confidence



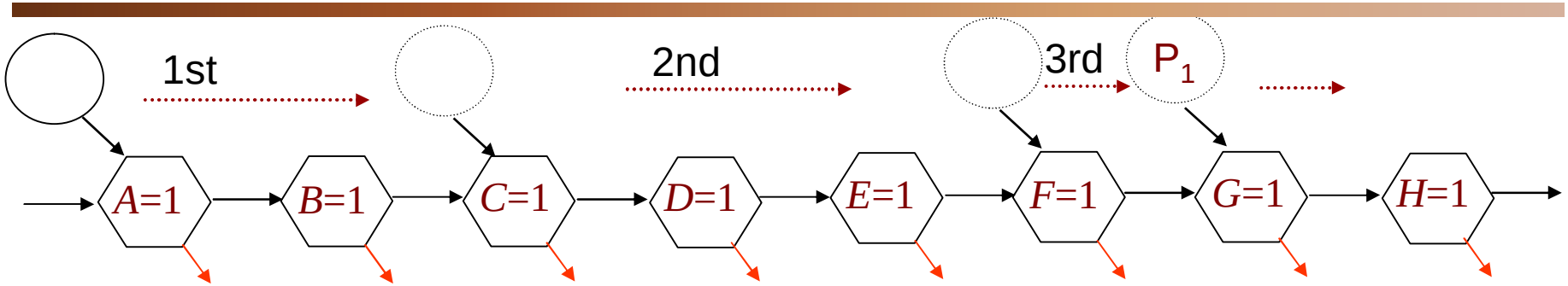
association rule	sup.	conf.
$A \Rightarrow B$	b / N	b / a
$A \Rightarrow B \wedge C$	c / N	c / a
$A \Rightarrow B \wedge C \wedge D$	d / N	d / a
$A \wedge B \Rightarrow C$	c / N	c / b
$A \wedge B \Rightarrow C \wedge D$	d / N	d / b
$A \wedge B \wedge C \Rightarrow D$	d / N	d / c

- χ^2 (chi-squared)



(ex.) $A \wedge B \Rightarrow C \wedge D$

	$C \wedge D$	$\neg(C \wedge D)$	
$A \wedge B$	**d**	b−d	**b**
$\neg(A \wedge B)$	δ −d	N-b- δ +d	N−b
	δ	N− δ	**N**



☐ GNP changes connection of Processing nodes in each generation

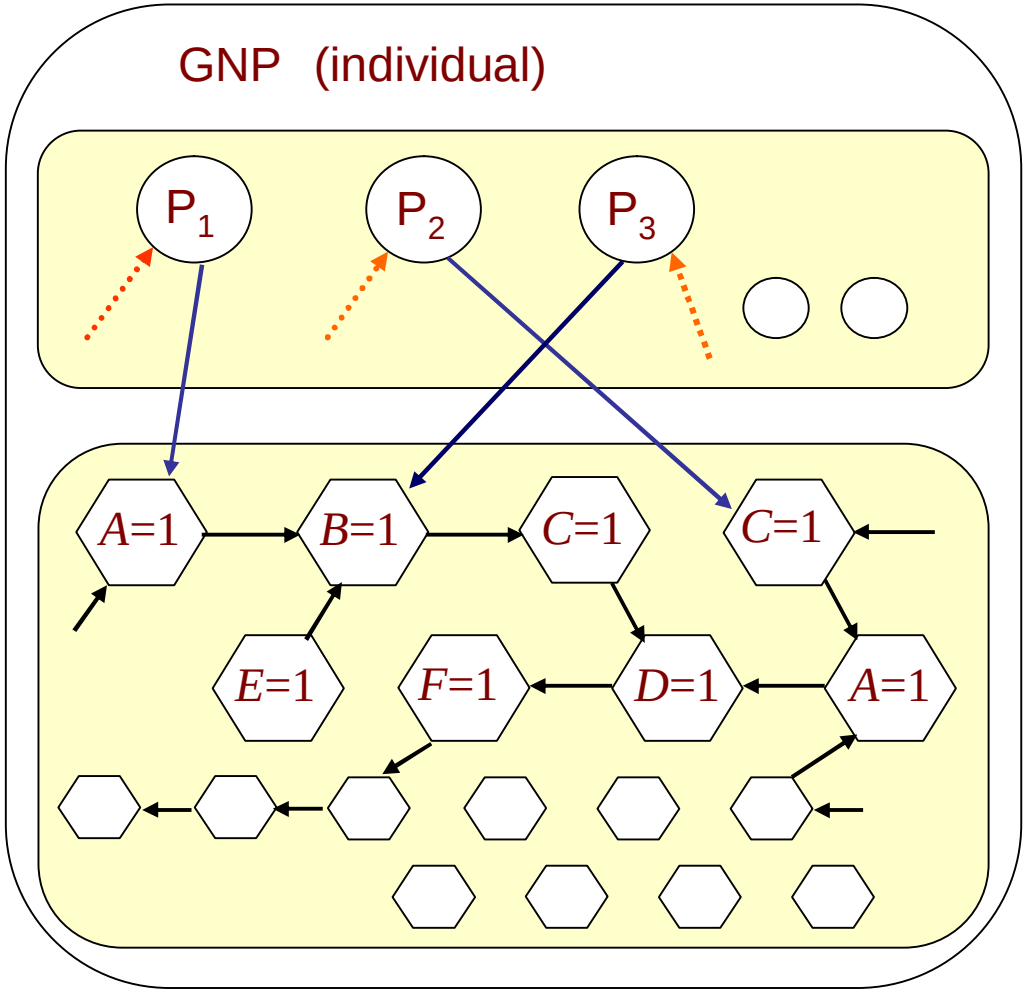
1) calculation of support

initial	1st	2nd	3rd
A	C	F	G
AB	CD	FG	GH
ABC	CDE	FGH	GHI
ABCD	CDEF	FGHI	GHIJ

2) calculation of chi-squared

1st	2nd	3rd
AB⇒C	CDE⇒F	F⇒G
AB⇒CD	CDE⇒FG	F⇒GH
AB⇒CDE	CDE⇒FGH	F⇒GHI
AB⇒CDEF	CDE⇒FGHI	F⇒GHIJ

Association Rule Mining Using GNP



: Processing node

connect to a judgement node

: Judgement node

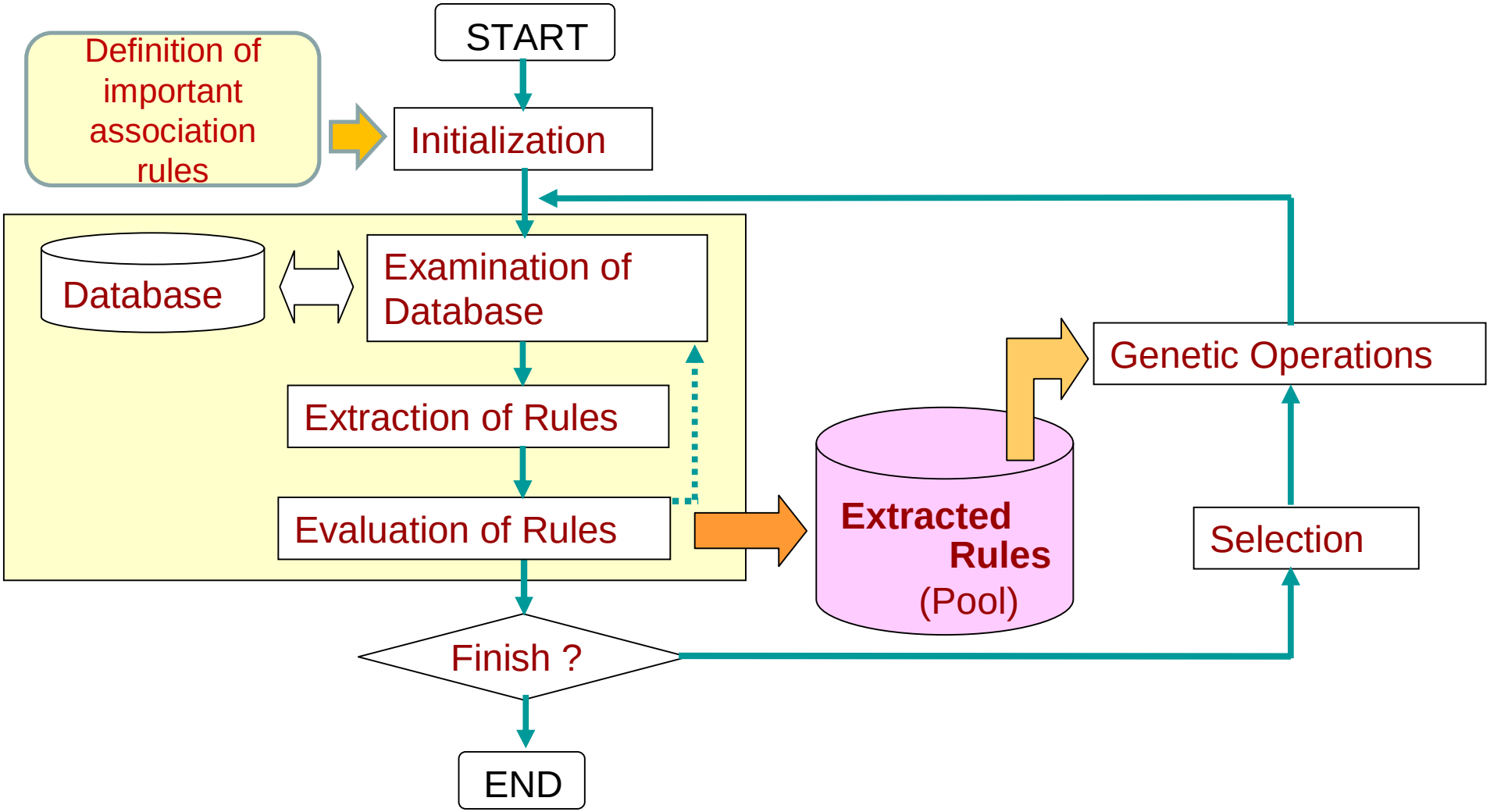
Yes

connect to a judgement node

No

connect to next numbered processing node

Association Rule Mining Using GNP



Fitness Function and Genetic Operators

- Definition of important association rules

$$\chi^2 > \chi^2_{min} \quad \text{and} \quad \text{support} \geqq sup_{min}$$

- Fitness evaluation function

$$Fitness = \sum_r \{\chi^2_r + 10(n_{ante}(r) - 1) + 10(n_{con}(r) - 1) + \alpha_{new}\}$$

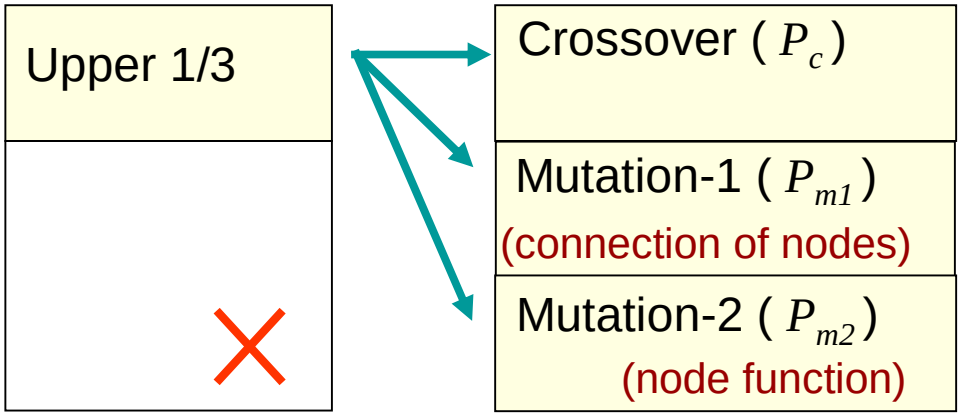
$n_{ante}(r)$: number of attributes of antecedent

$n_{con}(r)$: number of attributes of consequent

α_{new} : additional constant

- Selection and Genetic Operations

Ranked by fitness value



Use of Acquired Information

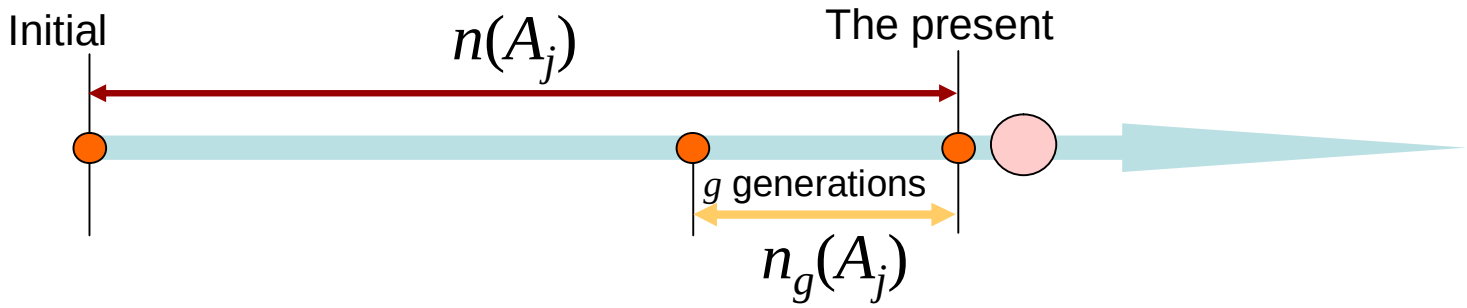
- Probability of selecting the Attribute A_j for judgement nodes (at Mutation-2)

$$P_j^{all} = \frac{n(A_j)+1}{\sum_k (n(A_k)+1)}$$

$n(A_j)$: frequency of the attribute A_j in the rules extracted in all generations

$$P_j^{g} = \frac{n_g(A_j)+1}{\sum_k (n_g(A_k)+1)}$$

$n_g(A_j)$: frequency of the attribute A_j in the rules extracted in the latest g generations



Experimental Results (1)

- Rule extraction from dense data set

  - Definition of

    " important association rules "

    $\chi^2 > 6.63, \quad sup_{min} = 0.1$

    $n_{ante}(r) + n_{con}(r) \geq 6$

    $n_{ante}(r) \leq 5, \quad n_{con}(r) \leq 5$

  - Synthetic database

    26 attributes (A_j , j =1, 2, …, 26)

    200 tuples

    $support(A_j=1)$ =0.7 (j =1, …, 5)

    $support(A_j=1)$ =0.5 (j =6, …, 26)

  - GNP

    Population size : 120

    Processing nodes : 10

    Judgement nodes : 78

    P_{m1} =1/3

    P_{m2} =1/5

    P_c =1/5

Experimental Results (1)

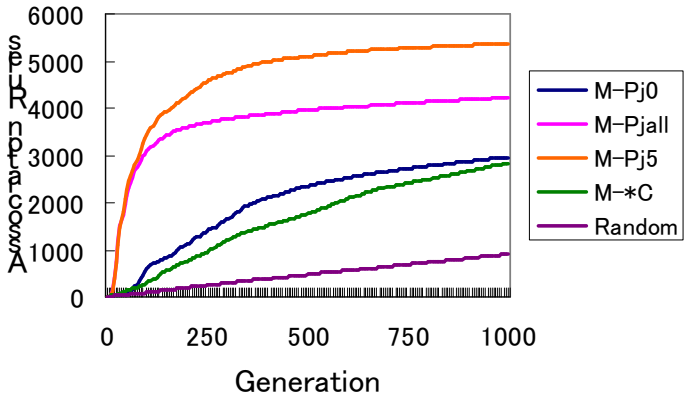


Fig.1 Averaged number of association rules over 10 trials in the pool

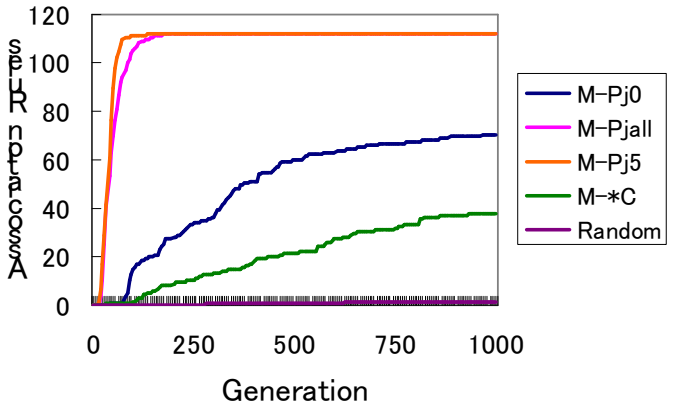


Fig.2 Averaged number of association rules over 10 trials in the pool

$(n_{ante}(r) + n_{con}(r) = 7)$

Using acquired information

P_j^5 $(g=5)$

P_j^{all}

Not using acquired information

P_j^0

Node functions are fixed

Random

Association Rule Mining Using GNP

- Extracts rules without identifying frequent itemsets used in Apriori-like mining methods
- Measures the significance of associations via the chi-squared test using GNP
- Stores extracted important association rules in a pool all together through generations
- Extracts important rules sufficiently enough for user's purpose in a short time.

Class Association Rule Mining

☐ Class Association Rule

$$(Z=1) \Leftrightarrow (A_j=1) \wedge \ldots \wedge (A_k=1)$$

TID	Z	A_1	A_2				A_n
1	1	0	1				0
2	0	1	0				1
3	1	1	0				0
…							
N	0	1	0				1

☐ Gene Analysis

$$(patient) \Leftrightarrow (SNP_j=G/G) \wedge \ldots \wedge (SNP_k=G/T)$$

SNP: Single Nucleotide Polymorphism

（遺伝子多型，一塩基多型）

--GATC**G**CAAT-----CAG**A**CCT--

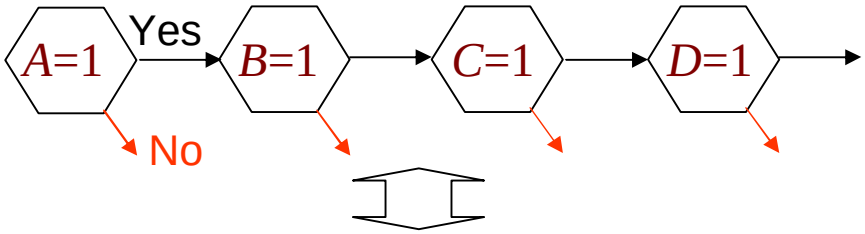
--GATC**T**CAAT-----CAG**T**CCT--

SNP1　　SNP2

TID	P	SNP1			SNP2						SNPn		
		G/G	G/T	T/T	A/A	A/T	T/T				G/G	G/C	C/C
1	1	0	0	1	1	0	0				0	1	0
2	0	0	1	0	1	0	0				0	0	1
…													
N	1	1	0	0	0	1	0				0	1	0

The Basic Ideas

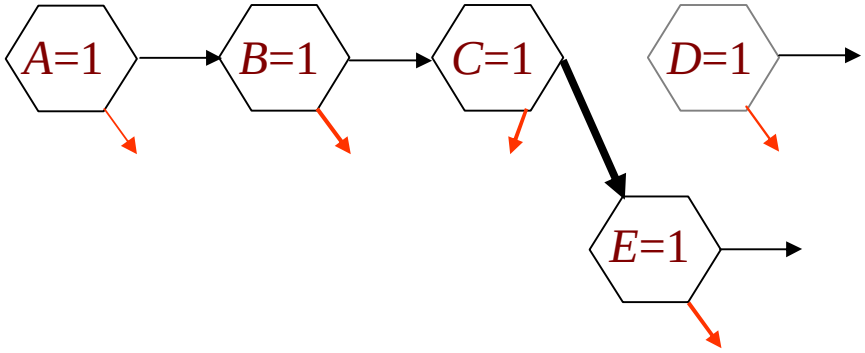
- Connect judgement nodes as antecedent of association rules



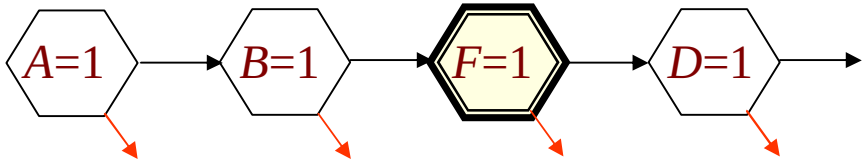
$$(A=1) \wedge (B=1) \wedge (C=1) \wedge (D=1) \Rightarrow (Z=1)$$

$$(A=1) \wedge (B=1) \wedge (C=1) \Rightarrow (Z=1)$$

- Obtain candidates by genetic operations



$$(A=1) \wedge (B=1) \wedge (C=1) \wedge \underline{(E=1)} \Rightarrow (Z=1)$$



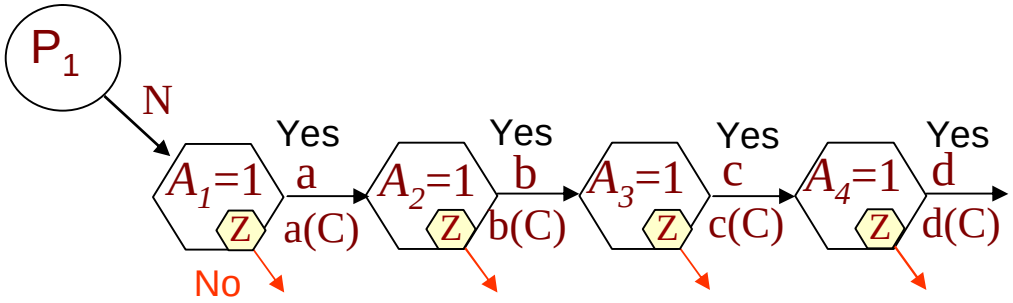
$$(A=1) \wedge (B=1) \wedge (C=1) \wedge \underline{(F=1)} \Rightarrow (Z=1)$$

Extraction of Class Association Rules

Class Association Rules

$$(A_j = 1) \wedge \ldots \wedge (A_k = 1) \Rightarrow (Z = C)$$

$$(C = 0, 1, \ldots, K)$$

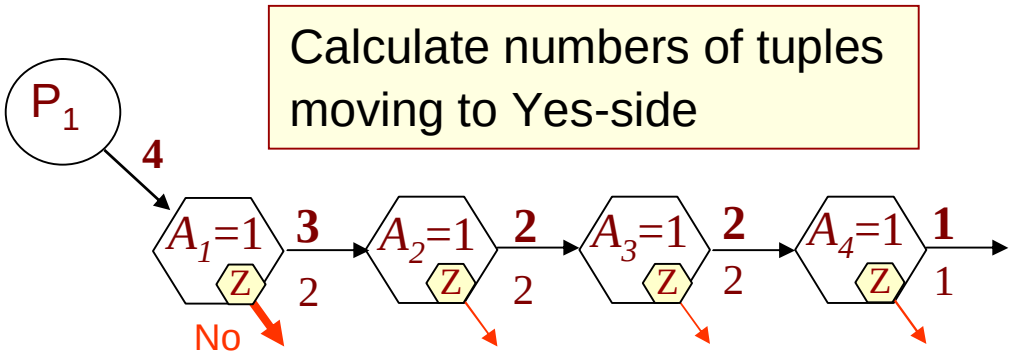


association rule	sup.	confi.
$A_1 \Rightarrow (Z = C)$	a(C) / N	a(C) / a
$A_1 \wedge A_2 \Rightarrow (Z = C)$	b(C) / N	b(C) / b
$A_1 \wedge A_2 \wedge A_3 \Rightarrow (Z = C)$	c(C) / N	c(C) / c
$A_1 \wedge A_2 \wedge A_3 \wedge A_4 \Rightarrow (Z = C)$	d(C) / N	d(C) / d

Extraction of Class Association Rules

Class Association Rules

$$(A_j = 1) \wedge \ldots \wedge (A_k = 1) \Rightarrow (Z = C)$$



	A_1	A_2	A_3	A_4	Z
1	1	0	1	0	0
2	1	1	1	1	1
3	1	1	1	0	1
4	0	1	0	1	1

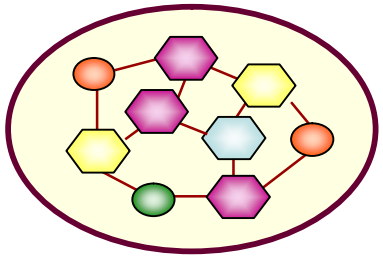
association rule	sup.	confi.
$A_1 \Rightarrow (Z = 1)$	2 / 4	2 / 3
$A_1 \wedge A_2 \Rightarrow (Z = 1)$	2 / 4	2 / 2
$A_1 \wedge A_2 \wedge A_3 \Rightarrow (Z = 1)$	2 / 4	2 / 2
$A_1 \wedge A_2 \wedge A_3 \wedge A_4 \Rightarrow (Z = 1)$	1 / 4	1 / 1

Experimental Results (2)

☐ Building a classifier using extracted rules

	A	B	C	D	…
00001	1	0	1	0	…
00002	0	1	1	1	…
00003	1	0	0	1	…
00004	0	1	0	0	…
…					

Training Data



GNP

Definition of important rules

$\chi^2_{min} = 6.63$

$\sup_{min} = 0.05$

$conf \geq \sup(Z = C)$

Class Association Rules

… …

$AGHWX \Rightarrow Z = 0$

$BJK \Rightarrow Z = 1$

… …

Dataset (UCI repository of ML database)

cleveland : 25 attributes, 297 tuples

breast-w : 18 attributes, 683 tuples

(Descretization of continulous attribute is done Entropy method)

Method 1

Method 2

	A	B	C	D	…
10001	1	0	1	0	…
10002	0	1	1	1	…
10003	1	0	0		
10004	0				
…					

Classification

10-fold cross validation

Building a classifier using extracted rules

Method 1

1) *total*(*C*) : compute the total number of rules satisfying $Z=C$ in the classifier (C=0, 1, 2, …, K)

2) *predict*(*C*) : compute the number of rules in the classifier, whose antecedent matches the items of the new data and satisfy $Z=C$ (C=0, 1, 2, …, K)

3) *score*(*C*)

$$score(C) = \frac{predict(C)}{total(C)}$$

(C=0, 1, 2, ..., K)

4) predict that the new data belong to the class having the highest *score*

Building a classifier using extracted rules

Method 2

1) *total*(*C*) : compute the total number of rules satisfying $Z=C$ in the classifier
(C=0, 1, 2, …, K)

2) calculate the distance $D_r(C)$ between the items of the new data and rule *r* which has the following *m* attributes in antecedent : $(A_{j1}=1)\wedge\ldots\wedge(A_{jm}=1)\Rightarrow(Z=C)$
Now, when the new data have *t* attributes which satisfies $A_{j_k}=1(1\le k\le m)$,
then, the new data have *m*-*t* attributes whose value is $A_{j_k}\neq 1$.

So, the distance $D_r(C)$ is calculated as $\max\left\{\frac{t-(m-t)}{m},0\right\}$

3) sum $D_r(C)$ to calculate *predict*(*C*) : $predict(C)=\sum_r D_r(C)$

4) s*core*(*C*)

$$score(C)=\frac{predict(C)}{total(C)}$$

(C=0, 1, 2, ..., K)

5) predict that the new data belong to the class having the highest *score*

Experimental Results (2)

Classification results versus **threshold of confidence** (Error rates (%))

confidence	Method 1			Method 2			CBA	c4.5
	all	0.8	0.9	all	0.8	0.9		
cleveland	17.7	17.7	18.3	16.3	17.3	17.7	16.7	18.2
breast-w	3.2	3.4	3.5	3.4	3.5	3.5	3.9	3.9

Classification results versus **number of rules** (Error rates (%))

Number of Rules	Method 1						Method 2					
	10	30	100	300	1000	all	10	30	100	300	1000	all
cleveland	30.3	19.3	18.3	18.0	17.3	17.7	27.3	19.7	18.0	17.0	17.0	16.3
breast-w	4.1	4.1	3.4	3.1	---	3.2	4.1	3.7	3.4	3.4	---	3.4

Experimental Results (2)

Classification results using **long** rules (Error rates (%))

Number of Rules	Method 1					Method 2				
	10	30	100	300	all	10	30	100	300	all
cleveland	34.7	28.3	22.0	21.0	19.7	18.3	18.0	19.7	19.0	19.0
breast-w	8.2	5.3	---	---	4.1	3.5	3.2	---	---	3.5

Averaged **run-time** versus number of extracted rules in the pool (sec)

Number of Rules	100	300	1000
cleveland (Z=1)	0.13	0.43	1.63
cleveland (Z=0)	0.14	0.41	1.60
beast-w (Z=1)	0.39	1.14	---
breast-w (Z=0)	0.35	1.38	---

Association rule mining between attribute families

$(A_j=1) \wedge \ldots \wedge (A_k=1) \Rightarrow (B_m=1) \wedge \ldots \wedge (B_n=1)$

$(B_m=1) \wedge \ldots \wedge (B_n=1) \Rightarrow (A_j=1) \wedge \ldots \wedge (A_k=1)$

TID	A_1	A_2	…	A_M	B_1	…	B_L
1	1	0		0	1		0
2	0	1		0	1		1
3	1	1		1	1		1
4	0	1		1	0		1
…						…	
N	0	1		1	0		1

Examples

(an itemset of genotype)

⇒ *(an itemset of environments or medical histories)*

(an itemset of environments or medical histories)

⇒ *(an itemset of genotype)*

☐ Our method can extract pairs of dependent sets of attributes between attribute families.

Apply to Genomics

- Association Study (Conventional method)

SNP-X

	normal	patient
G/G	250	235
G/T	500	505
T/T	250	260

SNP-Y

	normal	patient
G/G	750	30
G/T	150	580
T/T	100	390

SNP-Y is a candidate of the key SNP of the object

We need to examine SNP by SNP.
It is not easy to examine SNP combinations.

SNP (single nucleotide polymorphism)

--GATC**G**CAATC--
--GATC**T**CAATC--

Apply to Genomics (Association Study)

- Association Study (GNP-based method)

Extract important associations of SNPs and factors.

	$F_2F_4F_6$	$\neg(F_2F_4F_6)$
$S_1S_3S_7$	750	30
$\neg(S_1S_3S_7)$	250	970

SNP

Factor

TID	S_1	S_2	…	S_M	F_1	…	F_L
1	1	0		0	1		0
2	0	1		0	1		1
3	1	1		1	1		1
4	0	1		1	0		1
…						…	
N	0	1		1	0		1

Combine SNPs and factors automatically.

	set of factor	¬(set of factor)
set of SNP	750	30
¬(set of SNP)	250	970

$$(SNP\text{-}X = \mathrm{T/T}) \wedge \ldots \wedge (SNP\text{-}Z = \mathrm{G/G})$$
$$\Rightarrow (factor\text{-}j = 1) \wedge \ldots \wedge (factor\text{-}k = 1)$$