Discovering Association Patterns Based on Mutual Information

Bon K. Sy
Queens College/CUNY
Computer Science Department
bon@bunny.cs.qc.edu

Key points

- New concept for association patterns
  - support + interestingness/dependency
- Steps towards optimal association pattern discovery
  - Aprior property and mutual information measure
  - Probability model identification and inference
- What did we learn about proposed pattern approach?
  - Application of pattern approach to classification problems
  - Comparison: pattern approach, naïve Bayes & neural network

Agenda

- Meaning of association
- New concept on association pattern
- Association pattern discovery
- Pattern approach for classification problems
- Effectiveness and usefulness

Meaning of association

- What is an association pattern?
  - (A:0 B:1 C:0) where A,B,C are binary-valued variables
  - Pr(A:0 B:1 C:0) as a support measure

- Association rule A:0 -> B:1
  - Support: Pr(A:0 B:1)   Confidence: Pr(A:0|B:1)
  - Logic inference: Pr(A:0 -> B:1) ⇔ Pr(not(A:0) or B:1)
  - Conditional probability: Pr(B:1| A:0)

Deriving association rule is hard!

- Exponential number of combinations to make association rules from association patterns
  - Consider a pattern (A:0 B:1 C:0 D:1)
  - 46 possible rules in form of X -> Y even there are redundant rules; e.g. A:0 ->B:1 C:0 => A:0 B:1-> C:0

$$\sum_{i=1}^{n}\sum_{j=1}^{n-i}\binom{n}{i}\binom{n-i}{j}$$

Deriving association rule is hard!

- Spurious association
  - Suppose A:0 -> B:1 ⇔ if A:0 then B:1
    - How do we know it's not E:0 -> A:0 first, then E:0 -> B:1
- Interestingness or level of dependency
  - Pr(A:0 B:1) = 0.64, Pr(A:0) = Pr(B:1) = 0.8
    - Support measure = 0.64, confidence = 0.8
  - But A:0 and B:1 are independent of each other!

New concept for association patterns

- Criteria for significant association patterns:
  - Support measure
  - Interestingness/level of dependency
    - Mutual information measure in event pattern level:

$$Log_2 \frac{\Pr(A:0 \cap B:1)}{\Pr(A:0)\Pr(B:1)}$$

New concept for association patterns

- In two-variable case (Kullback)

$$E[MI(A,B)] = \sum \Pr(A,B) Log_2 \frac{\Pr(A \cap B)}{\Pr(A)\Pr(B)} \rightarrow \frac{\chi^2}{2N}$$

- Unfortunately statistical convergence does not behave well in high order patterns with multiple variables.

New concept for association patterns

- High order patterns with multiple variables

$$MI(x1, x2 \ldots xn) \rightarrow \left(\frac{1}{\Pr(x1, x2 \ldots xn)}\right)\left(\frac{\chi^2}{2N}\right)^{\left(\frac{\hat{E}}{E'}\right)^{\frac{o}{2}}}$$

where $MI(x1,x2 \ldots xn) = Log_2 Pr(x1\ x2 \ldots xn)/Pr(x1)Pr(x2) \ldots Pr(xn)$

N = sample population size

χ^2 = Pearson chi-square test statistic defined as $(oi - ei)^2/ei$

$\hat{E}$ = Expected entropy measure of estimated probability model

E' = Maximum possible entropy of estimated probability model

O = order of the association pattern (i.e., *n* in this case)

Association pattern discovery

- Discovering significant association patterns requires:
  - Joint probability information ***Pr(x1,x2,…xn)***
  - Marginal probability information ***Pr(x1), Pr(x2), … Pr(xn)***
  - Appropriate support threshold α related to population size ***N***
- Properties for significant association patterns:
  - Support measure ***Pr(x1,x2,…xn) >*** α %
  - -

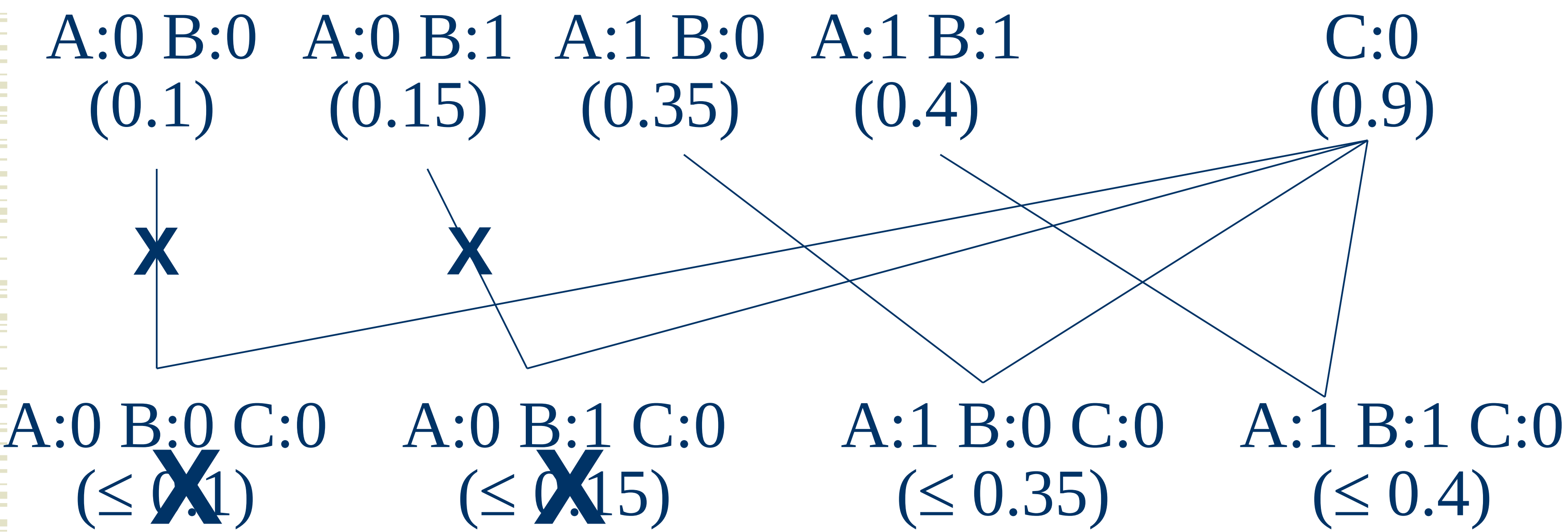
$$MI(x1, x2 \ldots xn) > \left(\frac{1}{\Pr(x1, x2 \ldots xn)}\right)\left(\frac{\chi^2}{2N}\right)^{\left(\frac{\hat{E}}{E'}\right)^{\frac{o}{2}}}$$

Association pattern discovery

- Bad news: Number of association patterns grows exponentially with respect to the order of the patterns.
- Good news: Properties for pruning
  - A priori property (Agrawal)
  - Mutual information convergence test

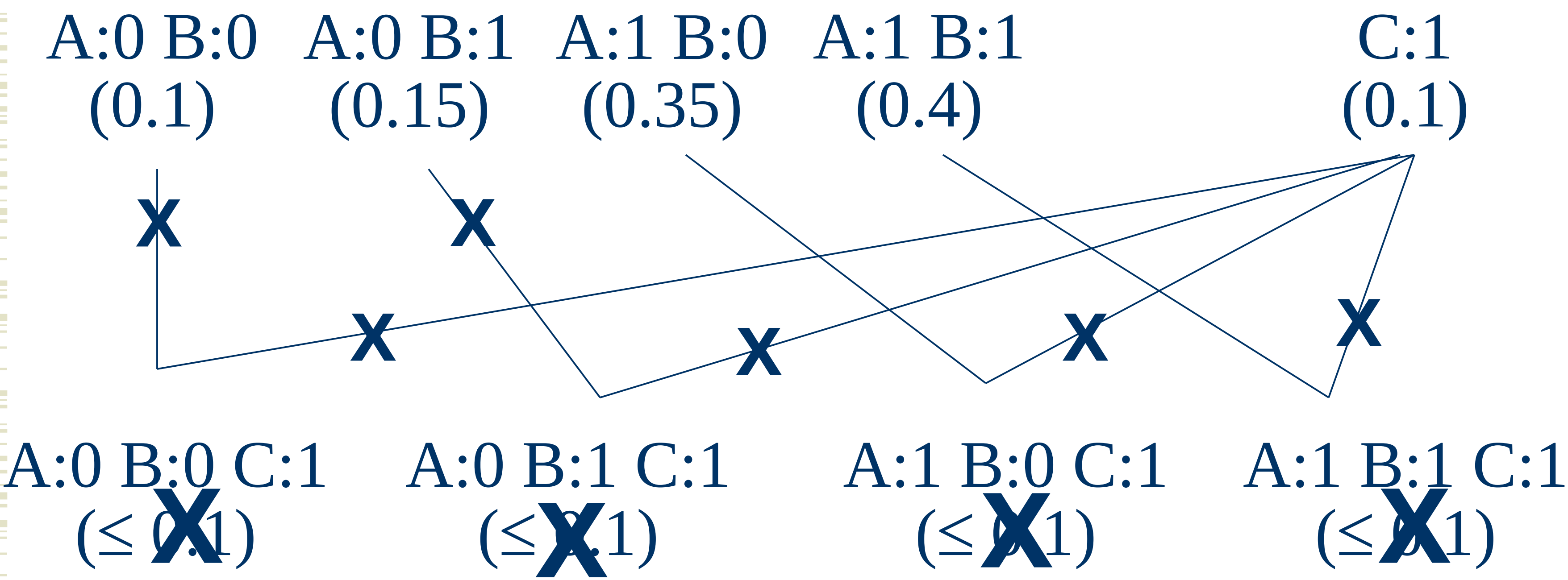
Association pattern discovery

- Pruning based on a prior property:



Association pattern discovery

- Pruning based on insufficient support:



Association pattern discovery

- Pruning based on mutual information criterion:

A:1 B:0
(0.35)

A:1 B:1
(0.4)

C:0
(0.9)

A:1 B:0 C:0
(≤ 0.35)

A:1 B:1 C:0
(≤ 0.4)

If *Pr(A:1,B:0,C:0)<0.30375*
Then *MI(A:1,B:0,C:0) < 0*

Association pattern discovery

- Trick on optimizing discovery process
  - Minimize counting related to high order pattern discovery based on low order pattern information:
    - Suppose
      - Pr(A:1,B:1,C:0) = 0.4
      - Pr(A:0,B:0,C:1) = 0.1
      - Pr(A:0,B:1,C:1) =0.1
      - Pr(A:1,B:0,C:1) =0.1
    - Then Pr(A:1, B:0,C:0) < 0.3

Association pattern discovery

- Key Concept (Sy 2001 and this paper):

Probabilistic inference on high order pattern information from existing information and that of low order patterns!

- Algorithm (in this paper) and implementation:
  - http://bonnet19.cs.qc.edu:7778/pls/rschdata/portal.login_dataMining
  - http://www.techsuite.net/kluwer/ (chapters 8 and 9)

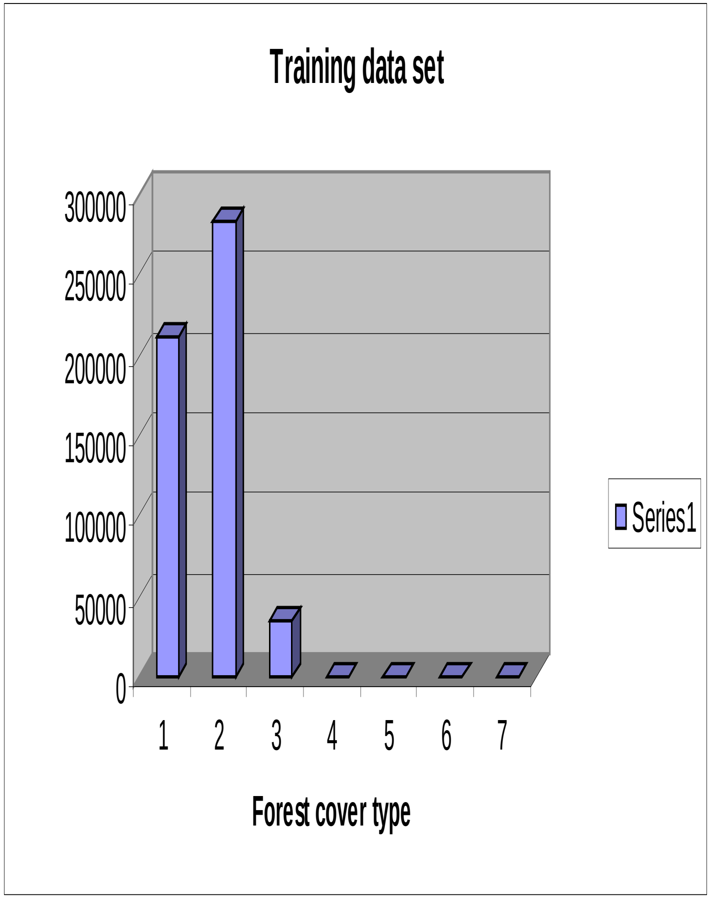
Pattern approach for classification (New! ☺)

- KDD forest cover type data set
  - 30x30 meter cells obtained from US Forest Service Region 2 Resource Information System (RIS) data.
    - http://kdd.ics.uci.edu/databases/covertype/covertype.data.html
    - http://kdd.ics.uci.edu/databases/covertype/covertype.task.html
- 581012 records; each with 54 attributes
  - 10 quantitative variables
  - 4 binary wilderness areas
  - 40 binary soil type variables (present/absent).

Pattern approach for classification (New! ☺)

- 7 types of tree coverage

TREE TYPE	COUNT
• Spruce/fir	211840
• Lodgepole pine	283301
• Ponderosa pine	35754
• Cottonwood/willow	2747
• Aspen	9493
• Douglas-fir	17367
• Krummholz	20510



Pattern approach for classification (New! ☺)

- Forest cover type classification problem:
  - Task 1 (Model attribute selection):
    - Identify few of the 40 binary-valued attributes about soil type information that covers sufficient data variability (measured by R-squared and residual).
  - Task 2 (Pattern discovery):
    - Identify statistically significant association patterns.
  - Task 3 (model identification):
    - Pattern-based classifier based on probability decision support models.
  - Task 4 (Comparative study):
    - Evaluate against naïve Bayes and neural network.

Pattern approach for classification (New! ☺)

- Experiment setup:
  - Training data: 11340
  - Validation data: 3780
  - Test data: 565892
- Task 1 (Model attribute selection):
  - Data set: Training data (11340 records of 40 attributes)
  - Least Square Trimmed robust and tree regressions
  - S-SPLUS 4.5, XP with 2G Hz CPU, 512M memory
  - Six soil type attributes as predictors
  - R-squared value: 0.9665. Tree: 21 levels, 96 terminals

Pattern approach for classification (New! ☺)

- Task 2 (Pattern discovery):
  - Data set: Training data (11340 records of 7 attributes)
  - Platform: Oracle 9.0.2 in Linux
  - Pattern discovery algorithm implemented as PL/SQL
  - Forest cover type as "response" variable (7 cases)
  - Any subset combination of value instantiation of the 6 predictors + response variable = association pattern
  - 64 (out of 3996) association patterns are significant.

Pattern approach for classification (New! ☺)

- Task 3 (Model identification):
  - Breakdown by forest cover types:

# of association patterns	Forest cover type
14	V55:1 (spruce/fir)
13	V55:2 (lodgepole pine)
6	V55:3 (ponderosa pine)
11	V55:4 (cottonwood/willow)
7	V55:5 (aspen)
7	V55:6 (Douglas-fir)
6	V55:7 (krummholz)

  - 7 probability decision models *M1*, … *M7*
    - Mi ⇔ Pr(V*24 V27 V44 V49 V52 V53 V55*$_i$')

Pattern approach for classification (New! ☺)

- Task 4 (Comparative study):
  - Pattern-based classifier:

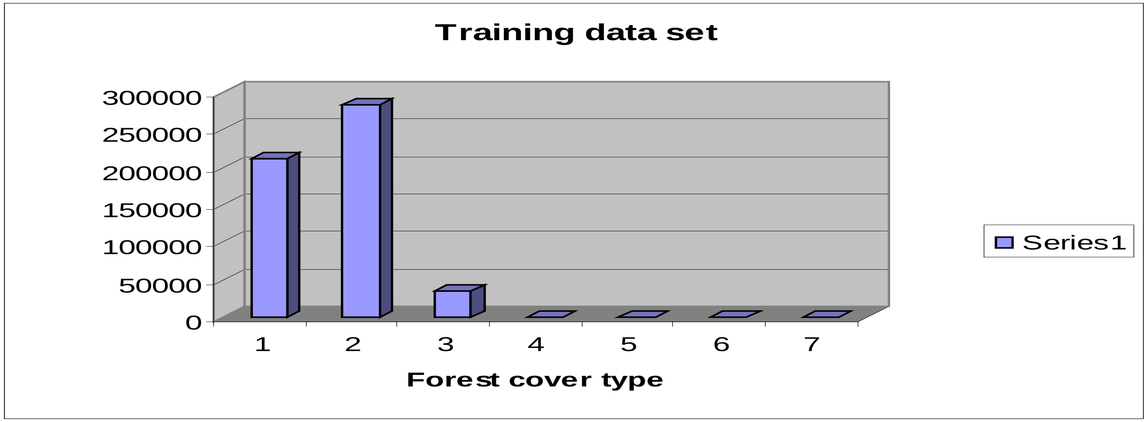
$$C(O) = ArgMax_{Mi}\ Pr(O|Mi) \text{x} W_{Mi}$$

$$W_{mi} = \textit{weighting factor} \propto \textit{population distribution}$$

  - 2 alternatives for comparison
    - Naïve Bayes and Neural network (Insightful Miner 1.0)
    - Neural network configuration:
      - Feed forward network
      - 1 fully connected hidden layer consisting of 10 nodes
      - Resilient propagation with a convergence tolerance = 0.0001, epochs = 50, and a learning rate = 0.01.

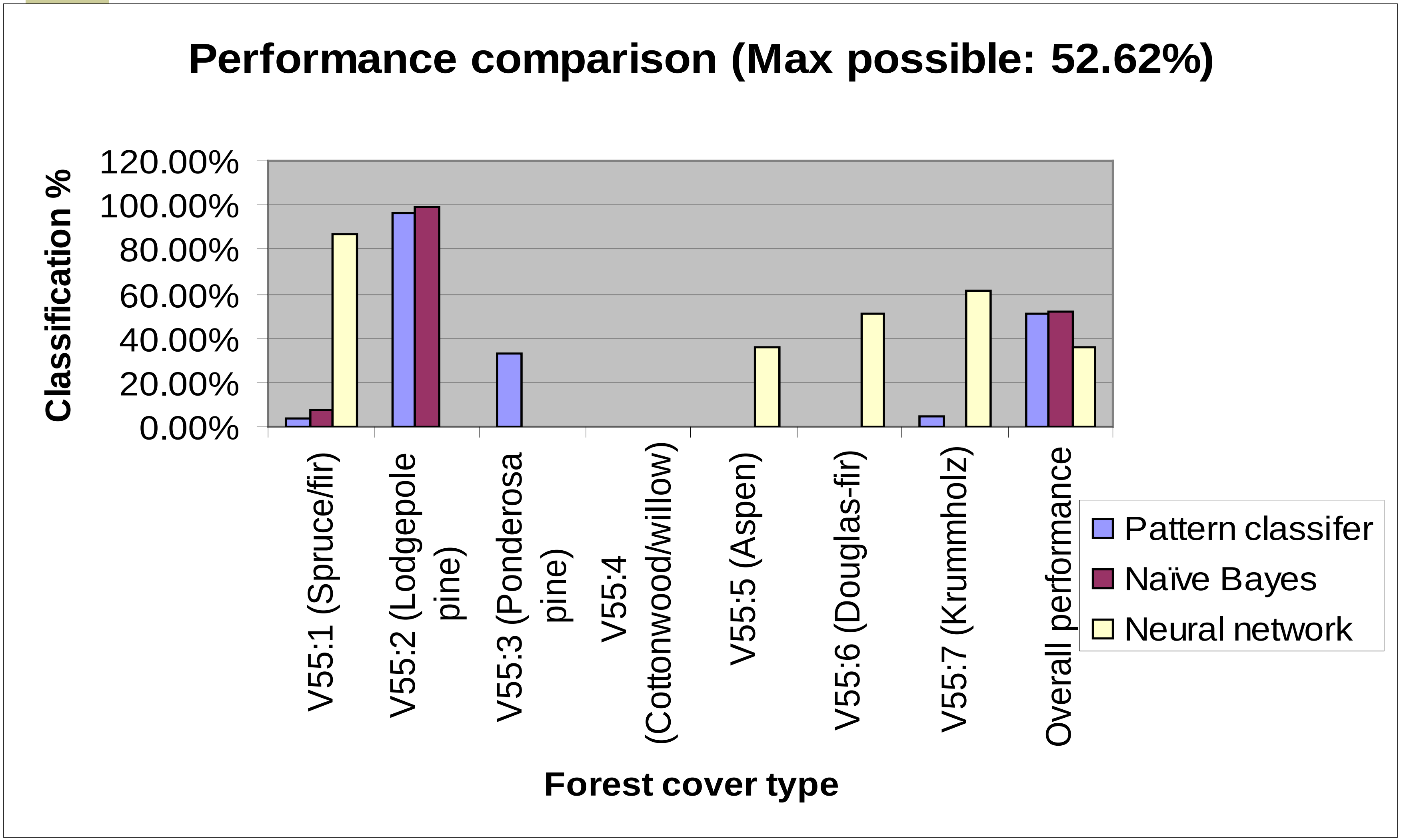
Effectiveness and usefulness (New! ☺)

- Same training data set is used in all three cases:
  - Pattern-based classifier; Naïve Bayes; Neural network



- Test data used for comparative study:
  - 565892 records of six soil type attributes.
  - Objective: Predict forest cover type.

Effectiveness and usefulness (New! ☺)



Test data Distribution By cover type

37.1% |49.7%|5.9% |0.1% |1.3% |2.7% |3.2%|

Effectiveness and usefulness (New! ☺)

- Observation: Neural network under-performed comparing to the two other approaches
- Hypothesis: Biased statistical distribution.
- Validation based on data set of even distribution
  - 3780 records of validation data (or 540 each)

Validation data	*Class 1*	*Class 2*	*Class 3*	*Class 4*	*Class 5*	*Class 6*	*Class 7*	*Overall performance*
Pattern classifier	*0%*	*0%*	*0%*	*91.70%*	*35.70%*	*52.60%*	*61.10%*	*34.40%*
Naïve Bayes	*0%*	*0%*	*0%*	*91.70%*	*35.70%*	*52.60%*	*61.10%*	*34.40%*
Neural net	*0%*	*0%*	*0%*	*91.70%*	*35.70%*	*52.60%*	*61.10%*	*34.40%*

Conclusion

- New concept of association patterns
- Algorithm for association pattern discovery
- Implementation as naïve PL/SQL inside Oracle data warehouse
- Comparative study that demonstrates competitive performance

Further information

- *Information-statistical data mining and Oracle basics for warehouse building*, with Arjun Gupta, ISBN 1-4020-7650-9, Kluwer academic publishers.
- Archived presentation slide:
  - http://www.techsuite.net/bonnet3/dm2003/DM2003.ppt
- Software for model identification:
  - http://www.techsuite.net/kluwer/ (chapter 9)

Further information

- Data warehouse system and association pattern discovery software
  - S-PLUS source code

    http://www.techsuite.net/kluwer/ (chapter 8)
  - Web accessible data warehouse:

    http://bonnet19.cs.qc.edu:7778/pls/rschdata/
  - Integrated environment for data warehouse and data mining:
    http://bonnet19.cs.qc.edu:7778/pls/rschdata/portal.login_dataMining

Further information

- Description of the data sets in the data warehouse:
  - Brief data definition for temperature, precipitation and water quality   E-community (case 5288)
  - Data dictionary, index locator, and table size for water quality data

    E-community (case 5265)
  - Description of the water quality data semantic meaning
    - E-community (case 5306)

Q and A

Bon K. Sy
Queens College/CUNY
Computer Science Department
bon@bunny.cs.qc.edu