CS522 Advanced Database Systems

Classification: Rule-based Classifiers

Chengyu Sun
California State University, Los Angeles

Rule-based Classification Example ...

- The Vertebrate dataset

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
human	warm	yes	no	no	mammals
python	cold	no	no	no	reptiles
salmon	cold	no	no	yes	fishes
whale	warm	yes	no	yes	mammals
frog	cold	no	no	sometimes	amphibians
komodo	cold	no	no	no	reptiles
bat	warm	yes	yes	no	mammals
pigeon	warm	no	yes	no	birds
cat	warm	yes	no	no	mammals
leopard shark	cold	yes	no	yes	fishes
turtle	cold	no	no	sometimes	reptiles
penguin	warm	no	no	sometimes	birds
porcupine	warm	yes	no	no	mammals
eel	cold	no	no	yes	fishes
salamander	cold	no	no	sometimes	amphibians
gila monster	cold	no	no	no	reptiles
platypus	warm	no	no	no	mammals
owl	warm	no	yes	no	birds
dolphin	warm	yes	no	yes	mammals
eagle	warm	no	yes	no	birds

... Rule-based Classification Example

- The Rules

r1: (Give Birth = no) ∧ (Can Fly = yes) → Birds
r2: (Give Birth = no) ∧ (Live in Water = yes) → Fishes
r3: (Give Birth = yes) ∧ (Blood Type = warm) → Mammals
r4: (Give Birth = no) ∧ (Can Fly = no) → Reptiles
r5: (Live in Water = sometimes) → Amphibians

©Tan, Steinbach, Kumar Introduction to Data Mining 2004

Terminology

Rule set: $R = (r_1 \vee r_2 \vee \ldots \vee r_k)$

Rule: r_i: $(\text{Condition}_i) \rightarrow c_i$

- $\text{Condition}_i = (A_1 \text{ op } v_1) \wedge (A_2 \text{ op } v_2) \wedge \ldots \wedge (A_k \text{ op } v_k)$
  - Rule antecedent, precondition
  - **Conjunct**: $(A_i \text{ op } v_i)$, $\text{op} \in \{=, \neq, <, >, \leq, \geq\}$
- c_i
  - Class label
  - Rule consequent

Coverage and Accuracy

- A rule `r` *covers* a record `x` if the precondition of `r` matches the attributes of `x`
  - A.K.A. `r` is *fired/triggered* by `x`
- Coverage(r) = |A| / |D|
  - |A|: # of records covered by r
- Accuracy(r) = |A∩y| / |A|
  - |A∩y|: # of records that satisfy both the antecedent and consequent of r
- Example
  - *coverage and accuracy of r3??*

How a Rule-based Classifier Works

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
lemur	warm	yes	no	no	?
turtle	cold	no	no	sometimes	?
dogfish shark	cold	yes	no	yes	?

- Lemur: ??
- Turtle: ??
- Dogfish shark: ??

©Tan, Steinbach, Kumar Introduction to Data Mining 2004

Two Properties of a Rule-based Classifier

- Exhaustive Rules
  - Every combination of the attribute values is covered by at least one rule
- Mutually Exclusive Rules
  - No two rules are triggered by the same record

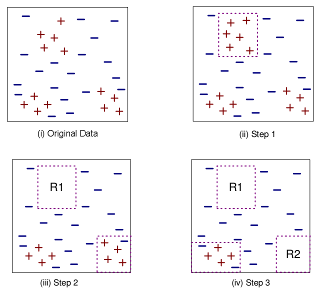
Make a Rule Set Exhaustive/Mutually Exclusive

- Default rule: $() \rightarrow c_d$
- Ordered rules
  - Quality-based ordering
  - Class-based ordering
- Unordered rules
  - Majority votes
    - Weighted by the rule's accuracy

Sequential Covering Algorithms

- Order the classes $\{c_1, c_2, ..., c_k\}$
- For each class c_i, $i<k$
  - Find the best rule r for c_i
  - Remove the records covered by r
  - Add r to the rule list
  - Repeat until the stop condition is met
- Add a default rule $() \rightarrow c_k$

Sequential Covering Example



Ordering Classes and Rules

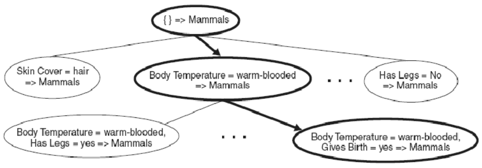
- Class ordering
  - Based on frequency
- Rule ordering
  - Based on classes
  - Based on quality of the rules

Rule Growing

- From general to specific
  - Start with $() \rightarrow c_i$
  - *Greedily* add one conjunct at a time
- From specific to general
  - Start with any positive record
  - *Greedily* remove one conjunct at a time

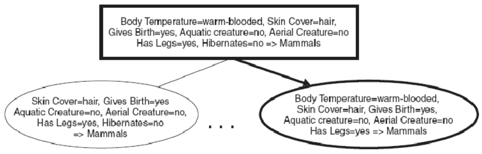
- Augmented by *beam search* with k best candidates

Rule Growing Example (a)



(a) General-to-specific

Rule Growing Example (b)



(b) Specific-to-general

Rule Evaluation

- Decide which conjunct should be added (or removed)

Rule Evaluation Example

- A training set contains 60 records in class c_1 and 100 records in class c_2
- Compare two rules
  - r_1: covers 50 c_1 and 5 c_2
  - r_2: covers 2 c_1 and 0 c_2

Rule Evaluation Measure (a)

Likelihood Ratio:

$$R(r) = 2\sum_{i=1}^{k} f_i \log(f_i / e_i)$$

f_i: observed # of class i records covered by r
e_i: expected # of class i records covered by r

Rule Evaluation Measure (b)

FOIL's information gain:

$$FGain(r) = n_c' \times (\log_2 \frac{n_c'}{n'} - \log_2 \frac{n_c}{n})$$

	# of records covered by r	# of correct records covered by r
Before rule growth	n	n_c
After rule growth	n'	n_c'

Stop Conditions

- Stop growing a rule
- Stop adding a rule for class c_i
  - Minimum Description Length (MDL)

Rule Pruning

- Similar to post-pruning of decision trees
- Remove a conjunct if the accuracy rate improves based on a validation set

Indirect Rule Extraction

- Using decision tree
  - Rule generation
  - *Exhaustive?? Mutually Exclusive??*
- Using association rule mining
  - Find association rules in the form of $\mathbf{A} \rightarrow c_i$
  - Select a subset of the rules to form a classifier
    - Sort the rules based on confidence, support, and length
    - Add to a rule list one at a time
    - Add a default rule

Readings

- Textbook Chapter 6.5