

CSE5334 DATA MINING

Lecture 5:
Classification (1)CSE4392/5334 Data Mining, Fall 2009
Department of Computer Science and Engineering, University of Texas at Arlington
Chengkai Li (Slides courtesy of Vipin Kumar)

Classification: Definition

- Given a collection of records (*training set*)
 - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
 - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 2

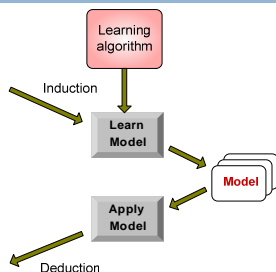
Illustrating Classification Task

Id	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	Yes
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Set

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

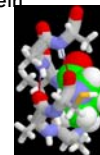
Test Set



Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 3

Examples of Classification Task

- Predicting tumor cells as benign or malignant
- Classifying credit card transactions as legitimate or fraudulent
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil
- Categorizing news stories as finance, weather, entertainment, sports, etc



Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 4

Classification vs. Prediction

- **Classification**
 - predicts categorical class labels
 - Most suited for nominal attributes
 - Less effective for ordinal attributes
- **Prediction**
 - models continuous-valued functions or ordinal attributes, i.e., predicts unknown or missing values
 - e.g., Linear regression

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 5

Supervised vs. Unsupervised Learning

- **Supervised learning (classification)**
 - Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations
 - New data is classified based on the training set
- **Unsupervised learning (clustering)**
 - The class labels of training data is unknown
 - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 6

Classification Techniques

- Decision Tree based Methods
- Rule-based Methods
- Nearest-Neighbor Classifiers
- Naïve Bayes Classifiers and Bayesian Belief Networks
- Neural Networks
- Support Vector Machines

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 7

Example of a Decision Tree

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data Model: Decision Tree

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 8

Another Example of Decision Tree

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

There could be more than one tree that fits the same data!

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 9

Decision Tree Classification Task

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Set

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	45K	?
No	Married	80K	?

Test Set

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 10

Apply Model to Test Data

Start from the root of tree.

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 11

Apply Model to Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 12

Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 13

Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 14

Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 15

Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 16

Decision Tree Classification Task

Id	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Set

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	55K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

Test Set

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 17

Decision Tree Induction

- Large search space
 - Exponential size, with respect to the set of attributes
 - Finding the optimal decision tree is computationally infeasible
- Efficient algorithm for accurate suboptimal decision tree
 - Greedy strategy
 - Grow the tree by making locally optimally decisions in selecting the attributes

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 18

Decision Tree Induction

- Many Algorithms:
 - Hunt's Algorithm (one of the earliest, basis of others)
 - CART
 - ID3, C4.5
 - SLIQ,SPRINT

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 19

General Structure of Hunt's Algorithm

- Let D_t be the set of training records that reach a node t
- General Procedure:
 - If D_t contains records that belong to the same class y_t , then t is a leaf node labeled as y_t .
 - If D_t is an empty set, then t is a leaf node labeled by the default class, y_d .
 - If D_t contains records that belong to more than one class, use an **attribute test** to split the data into smaller subsets. Recursively apply the procedure to each subset.

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

D_t

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 20

Hunt's Algorithm

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 21

Tree Induction

- Greedy strategy.
 - Split the records based on an attribute test that optimizes certain criterion.
- Issues
 - Determine how to split the records
 - How to specify the attribute test condition?
 - How to determine the best split?
 - Determine when to stop splitting

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 22

Tree Induction

- Greedy strategy.
 - Split the records based on an attribute test that optimizes certain criterion.
- Issues
 - Determine how to split the records
 - How to specify the attribute test condition?
 - How to determine the best split?
 - Determine when to stop splitting


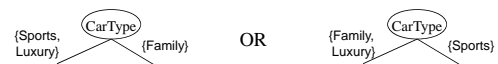
Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 23

How to Specify Test Condition?

- Depends on attribute types
 - Nominal
 - Ordinal
 - Continuous
- Depends on number of ways to split
 - 2-way split
 - Multi-way split


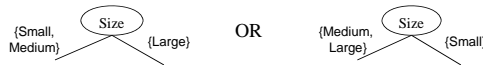

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 24

Splitting Based on Nominal Attributes

- Multi-way split: Use as many partitions as distinct values.
 
- Binary split: Divides values into two subsets. Need to find optimal partitioning.
 

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 25

Splitting Based on Ordinal Attributes

- Multi-way split: Use as many partitions as distinct values.
 
- Binary split: Divides values into two subsets. Need to find optimal partitioning.
 
- What about this split?
 

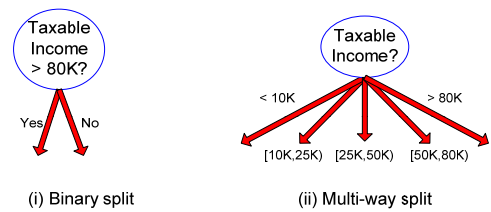
Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 26

Splitting Based on Continuous Attributes

- Different ways of handling
 - Discretization to form an ordinal categorical attribute
 - Static – discretize once at the beginning
 - Dynamic – ranges can be found by equal interval bucketing, equal frequency bucketing (percentiles), or clustering.
 - Binary Decision: $(A < v)$ or $(A \geq v)$
 - consider all possible splits and finds the best cut
 - can be more computational intensive

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 27

Splitting Based on Continuous Attributes



(i) Binary split (ii) Multi-way split

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 28

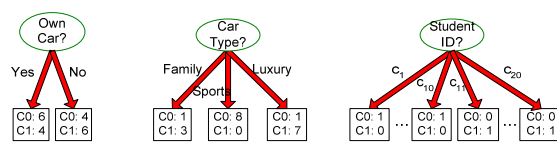
Tree Induction

- Greedy strategy.
 - Split the records based on an attribute test that optimizes certain criterion.
- Issues
 - Determine how to split the records
 - How to specify the attribute test condition?
 - How to determine the best split?
 - Determine when to stop splitting

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 29

How to determine the Best Split

Before Splitting: 10 records of class 0, 10 records of class 1



Which test condition is the best?

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 30

How to determine the Best Split

- Greedy approach:
 - Nodes with **homogeneous** class distribution are preferred
- Need a measure of node impurity:

C0: 5
C1: 5

Non-homogeneous,
High degree of impurity

C0: 9
C1: 1

Homogeneous,
Low degree of impurity

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 31

Measures of Node Impurity

- Gini Index
- Entropy
- Misclassification error

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 32

How to Find the Best Split

Before Splitting:

C0	N00
C1	N01

 → M0

A?

Yes → Node N1
No → Node N2

C0	N10
C1	N11

M1 M2

M12

B?

Yes → Node N3
No → Node N4

C0	N30
C1	N31

M3 M4

M34

$M0 - M12$ vs $M0 - M34$

(Information gain, if Entropy is used as M)

33

Measure of Impurity: GINI

- Gini Index for a given node t :

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

(NOTE: $p(j|t)$ is the relative frequency of class j at node t .)

 - Maximum $(1 - 1/n_t)$ when records are equally distributed among all classes, implying least interesting information
 - Minimum (0.0) when all records belong to one class, implying most interesting information

C1	0
C2	6
GINI	0.000

C1	1
C2	5
GINI	0.278

C1	2
C2	4
GINI	0.444

C1	3
C2	3
GINI	0.500

34

Examples for computing GINI

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

C1	0
C2	6

$P(C1) = 0/6 = 0$ $P(C2) = 6/6 = 1$
 $Gini = 1 - P(C1)^2 - P(C2)^2 = 1 - 0 - 1 = 0$

C1	1
C2	5

$P(C1) = 1/6$ $P(C2) = 5/6$
 $Gini = 1 - (1/6)^2 - (5/6)^2 = 0.278$

C1	2
C2	4

$P(C1) = 2/6$ $P(C2) = 4/6$
 $Gini = 1 - (2/6)^2 - (4/6)^2 = 0.444$

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 35

Splitting Based on GINI

- Used in CART, SLIQ, SPRINT.
- When a node p is split into k partitions (children), the quality of split is computed as,

$$GINI_{split} = \sum_{i=1}^k \frac{n_i}{n} GINI(i)$$

where, n_i = number of records at child i ,
 n = number of records at node p .

How to Find the Best Split

Before Splitting:

C0	N00
C1	N01

 → M0

$GINI_{split} = \sum_{i=1}^k \frac{n_i}{n} GINI(i)$

A?
 Yes → Node N1
 No → Node N2

B?
 Yes → Node N3
 No → Node N4

Node N1:

C0	N10
C1	N11

 → M1

Node N2:

C0	N20
C1	N21

 → M2

Node N3:

C0	N30
C1	N31

 → M3

Node N4:

C0	N40
C1	N41

 → M4

M12 (M1, M2)
 M34 (M3, M4)

gain (Information gain, if Entropy is used as M)
 M0 - M12 vs M0 - M34

Binary Attributes: Computing GINI Index

- Splits into two partitions
- Effect of Weighing partitions:
 - Larger and Purer Partitions are sought for.

B?
 Yes → Node N1
 No → Node N2

Parent:

	Parent
C1	6
C2	6

 Gini = 0.500

Node N1:

	N1	N2
C1	5	1
C2	2	4

 Gini(N1) = $1 - (5/7)^2 - (2/7)^2 = 0.408$
 Gini(N2) = $1 - (1/5)^2 - (4/5)^2 = 0.32$
 Gini(Children) = $7/12 * 0.408 + 5/12 * 0.32 = 0.371$

Categorical Attributes: Computing Gini Index

- For each distinct value, gather counts for each class in the dataset
- Use the count matrix to make decisions

Multi-way split

	CarType		
	Family	Sports	Luxury
C1	1	8	1
C2	3	0	7
Gini	0.163		

Two-way split (find best partition of values)

	CarType (Sports, Luxury)	
	Family	Luxury
C1	9	1
C2	7	3
Gini	0.468	

	CarType (Sports, Family)	
	Luxury	Family
C1	8	2
C2	0	10
Gini	0.167	

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 39

Continuous Attributes: Computing Gini Index

- Use Binary Decisions based on one value
- Several Choices for the splitting value
 - Number of possible splitting values = Number of distinct values
- Each splitting value has a count matrix associated with it
 - Class counts in each of the partitions, A < v and A ≥ v
- Simple method to choose best v
 - For each v, scan the database to gather count matrix and compute its Gini index
 - Computationally Inefficient! Repetition of work.

Trd	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Taxable Income > 80K?
 Yes / No

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 40

Continuous Attributes: Computing Gini Index...

- For efficient computation: for each attribute,
 - Sort the attribute on values
 - Linearly scan these values, each time updating the count matrix and computing gini index
 - Choose the split position that has the least gini index

	Taxable Income																	
	60	70	75	85	90	95	100	120	125	220								
Sorted Values	55	65	72	80	87	92	97	110	122	172	230							
Split Positions	<<	>	<<	>	<<	>	<<	>	<<	>	<<	>						
Yes	0	3	0	3	0	3	1	2	1	3	0	3	0					
No	0	7	1	6	2	5	3	4	3	4	4	3	5	2	6	1	7	0
Gini	0.420	0.400	0.375	0.343	0.417	0.400	0.390	0.343	0.375	0.400	0.420							

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 41

Alternative Splitting Criteria based on INFO

- Entropy at a given node t:

$$Entropy(t) = -\sum_j p(j|t) \log p(j|t)$$

(NOTE: p(j|t) is the relative frequency of class j at node t.)

 - Measures homogeneity of a node.
 - Maximum (log n_c) when records are equally distributed among all classes implying least information
 - Minimum (0.0) when all records belong to one class, implying most information
 - Entropy based computations are similar to the GINI index computations

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 42

Examples for computing Entropy

$$Entropy(t) = -\sum_j p(j|t) \log_2 p(j|t)$$

C1	0	P(C1) = 0/6 = 0	P(C2) = 6/6 = 1
C2	6	Entropy = -0 log 0 - 1 log 1 = -0 - 0 = 0	

C1	1	P(C1) = 1/6	P(C2) = 5/6
C2	5	Entropy = -(1/6) log ₂ (1/6) - (5/6) log ₂ (5/6) = 0.65	

C1	2	P(C1) = 2/6	P(C2) = 4/6
C2	4	Entropy = -(2/6) log ₂ (2/6) - (4/6) log ₂ (4/6) = 0.92	

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 43

Splitting Based on INFO...

Information Gain:

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^k \frac{n_i}{n} Entropy(i) \right)$$

Parent Node, p is split into k partitions;
n_i is number of records in partition i

- Measures Reduction in Entropy achieved because of the split. Choose the split that achieves most reduction (maximizes GAIN)
- Used in ID3 and C4.5
- Disadvantage: Tends to prefer splits that result in large number of partitions, each being small but pure.

Splitting Based on INFO...

Gain Ratio:

$$GainRATIO_{split} = \frac{GAIN_{split}}{SplitINFO} \quad SplitINFO = -\sum_{i=1}^k \frac{n_i}{n} \log \frac{n_i}{n}$$

Parent Node, p is split into k partitions
n_i is the number of records in partition i

- Adjusts Information Gain by the entropy of the partitioning (SplitINFO). Higher entropy partitioning (large number of small partitions) is penalized!
- Used in C4.5
- Designed to overcome the disadvantage of Information Gain

Splitting Criteria based on Classification Error

Classification error at a node t :

$$Error(t) = 1 - \max_i P(i|t)$$

- Measures misclassification error made by a node.
 - Maximum (1 - 1/n_i) when records are equally distributed among all classes, implying least interesting information
 - Minimum (0.0) when all records belong to one class, implying most interesting information

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 46

Examples for Computing Error

$$Error(t) = 1 - \max_i P(i|t)$$

C1	0	P(C1) = 0/6 = 0	P(C2) = 6/6 = 1
C2	6	Error = 1 - max (0, 1) = 1 - 1 = 0	

C1	1	P(C1) = 1/6	P(C2) = 5/6
C2	5	Error = 1 - max (1/6, 5/6) = 1 - 5/6 = 1/6	

C1	2	P(C1) = 2/6	P(C2) = 4/6
C2	4	Error = 1 - max (2/6, 4/6) = 1 - 4/6 = 1/3	

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 47

Comparison among Splitting Criteria

For a 2-class problem:
(p is the fraction of records belonging to one of the two classes.)

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 48

Tree Induction

- Greedy strategy.
 - ▣ Split the records based on an attribute test that optimizes certain criterion.
- Issues
 - ▣ Determine how to split the records
 - How to specify the attribute test condition?
 - How to determine the best split?
 - ▣ Determine when to stop splitting

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 49

Stopping Criteria for Tree Induction

- Stop expanding a node when all the records belong to the same class
- Stop expanding a node when all the records have same (or similar) attribute values
 - ▣ What to do? majority voting
- Early termination (to be discussed later)

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 50

Decision Tree Based Classification

- Advantages:
 - ▣ Inexpensive to construct
 - ▣ Extremely fast at classifying unknown records
 - ▣ Easy to interpret for small-sized trees
 - ▣ Accuracy is comparable to other classification techniques for many simple data sets

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 51

Example: C4.5

- Simple depth-first construction.
- Uses Information Gain
- Sorts Continuous Attributes at each node.
- Needs entire data to fit in memory.
- Unsuitable for Large Datasets.
 - ▣ Needs out-of-core sorting.
- You can download the software from:
 - ▣ <http://www.cse.unsw.edu.au/~quinlan/c4.5r8.tar.gz>

Lecture 5: Classification (I) CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 52