Knowledge Discovery and Data Mining

Unit # 4

Acknowledgement

- Most of the slides in this presentation are taken from course slides provided by
  - Han and Kimber (Data Mining Concepts and Techniques) and
  - Tan, Steinbach and Kumar (Introduction to Data Mining)

Background

- ID3 (Iterative Dichotomiser 3)
  - published in 1986 but proposed in 1983
  - Only works on non-continuous (discrete) attributes
  - Uses Information Gain/Entropy as the splitting rule
- CART
  - Published in 1984
  - Uses Gini Index as the splitting rule
  - Binary trees
- C4.5
  - Extension of ID3 and published in 1993
  - Works on continuous attributes
  - Uses modified Gain/Entropy metric as the splitting rule to defy advantage to variables having multiple states

Handling of Multi-state Variable

- The way both Gini Index and Entropy are presented, they become biased to variables having multiple states.
- To over this, the following approach was recommended (in C4.5 using Entropy but can be generalized to Gini Index as well).
  - Gain = SR (D) – SR_A (D)
  - Where SR = splitting rule metric
  - D = class variable
  - A = an attribute on which the splitting rule is conditioned

Buy Computer Example

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

SplitInfo

- Gini (buy) = 0.46
  - Gini $_{Age}$ (buy) = 0.34 : Gain = 0.12
  - Gini $_{inc}$ (buy) = 0.44 : Gain = 0.02
  - Gini $_{std}$ (buy) = 0.37 : Gain = 0.09
  - Gini $_{rat}$ (buy) = 0.43 : Gain = 0.03
- SplitInfo = unconditional splitting rules on the variables. If one is using Gini then it becomes
  - Splitinfo (age) = Gini (age) = 0.66
  - Splitinfo (inc) = Gini (inc) = 0.65
  - Splitinfo (std) = Gini (std) = 0.5
  - Splitinfo (rat) = Gini (rat) = 0.49

Gain_ratio

- To obtain gain ratio, we divide gain by splitinfo
  - Gain_ratio (age) = 0.12 / 0.66 = 0.18 (0.175)
  - Gain_ratio (inc) = 0.02 / 0.65 = 0.03
  - Gain_ratio (std) = 0.09 / 0.5 = 0.18 (0.184)
  - Gain_ratio (rat) = 0.03 / 0.49 = 0.06
- A similar computation would have been done if we were using Entropy or even Miclassification Error

Categorical Attributes: Computing GINI Index

- From a historical perspective, Gini Index always created a binary tree.
- As a result, in case of multiple values, it merged them together to find the best binary split
- For each distinct value, gather counts for each class in the dataset

Multi-way split

	CarType		
	Family	Sports	Luxury
C1	1	2	1
C2	4	1	1
Gini	0.393		

Two-way split
(find best partition of values)

	CarType	
	{Sports, Luxury}	{Family}
C1	3	1
C2	2	4
Gini	0.400	

	CarType	
	{Sports}	{Family, Luxury}
C1	2	2
C2	1	5
Gini	0.419	

Feature Discetization

- Unsupervised Discretization
  - Used in Clustering
- Supervised Discretization
  - Used in Classification

Unsupervised Feature Discretization Techniques

- The task of feature discretization techniques is to discretize the values of continuous features into a small number of intervals, where each interval is mapped to a discrete symbol.
- Suppose the set of values for a given feature are {3, 2, 1, 5, 4, 3, 1, 7, 5, 3}. After sorting, these values can be placed into three bins
  - {1, 1, 2, 3, 3, 3, 4, 5, 5, 7}

Value Reduction

- One of the main problems of the previous method is to find the best cutoffs for bins.
- The value-reduction problem can be stated as an optimization problem in the selection of k bins: given the number of bins k, distribute the values in the bins to minimize the average distance of a value from its bin mean or median.
- The distance is usually measured as the squared distance for a bin mean and as the absolute distance for a bin median.

Value Reduction – A Heuristic Algorithm

- Sort all values for a given feature.
- Assign approximately equal number of sorted adjacent values (vi) to each bin, where the number of bins is given in advance.
- Move a border element vi from one bin to the next (or previous) when that reduces the global distance error (ER) (the sum of all distances from each vi to the mean or mode of its assigned bin).

Working of the Algorithm

- The set of values for a feature f is {5, 1, 8, 2, 2, 9, 2, 1, 8, 6}.
- Split them into three bins (k = 3), where the bins will be represented by their modes.
- Initial bins are {1, 1, 2 2, 2, 5 6, 8, 8, 9}
- Modes for the three bins are {1, 2, 8}. The error, ER, is 0+0+1+0+0+3+2+0+0+1=7
- After moving two elements from BIN2 into BIN1 and one element from BIN3 to BIN2 in the next three iterations, the final distribution of elements are {1, 1, 2, 2, 2, 5, 6, 8, 8, 9}
- The total minimized error, ER, is 4.

Value Reduction Exercise

- Perform Bin-based values reduction with the best cutoffs for the following:
  - The feature Attribute 2 (in slide # 35, Unit # 3) using mean values as representatives for two bins.
  - Repeat the same exercise for three bins

Supervised Feature Discretization Technique: Chimerge

- Chimerge is one automated discretization algorithm that analyzes the quality of multiple intervals for a given feature by using χ^2 statistics.
- The algorithm consists of three basic steps:
  - Sort the data for the given feature in ascending order.
  - Define initial intervals so that every value is in a separate interval.
  - Repeat until no χ^2 of any two adjacent intervals is less then threshold value.

Chimerge Formula

- $\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} (A_{ij} - E_{ij})^2 / E_{ij}$
  - K = number of classes
  - Aij = number of instances in the i-th interval, j-th class
  - Eij = expected frequency of Aij, computed as (Ri . Cj)/N
  - Ri = number of instances in the i-th interval
  - Cj = number of instances in the j-th class
  - N = total number of instances
- If either Ri or Cj is 0, Eij is set to a small value.

	Class 1	Class 2	
Interval I	A_{11}	A_{12}	R_1
Interval 2	A_{21}	A_{22}	R_2
Σ	C_1	C_2	Σ

Chimerge Example

- For this example, interval points for feature F are 0, 2, 5, 7.5, 8.5, 10, etc.

	Class 1	Class 2	
[7.5, 8.5]	1	0	1
[8.5, 10]	1	0	1
Σ	2	0	2

- $\chi^2 = (1-1)^2/1 + (0-0.1)^2/0.1 + (1-1)^2/1 + (0-0.1)^2/0.1 = 0.2$
- For the degree of freedom d=1, $\chi^2 = 0.2 < 2.706$ (for $\alpha = 0.1$). We can conclude that there are no significant differences in relative class frequencies and that the selected intervals can be merged.

F	K
1	1
3	2
7	1
8	1
9	1
11	2
23	2
37	1
39	2
45	1
46	1
59	1

Chimerge Example (Cont'd)

- After several iterations we won't be able to merge intervals further.

	Class 1	Class 2	
[0, 10]	4	1	5
[10, 42]	1	3	4
Σ	5	4	9

- $\chi^2 = (4-2.78)^2/2.78 + (1-2.22)^2/2.22 + (1-2.22)^2/2.22 + (3-1.78)^2/1.78 = 2.72$
- For the degree of freedom d=1, $\chi^2 = 2.72 > 2.706$ (for $\alpha = 0.1$). The conclusion is that significant differences exist between two intervals and merging is not recommended.

ChiMerge Exercise

- Apply the ChiMerge technique to reduce the number of values for numeric attributes (Slide # 30, Unit # 2)
  - Reduce the number of numeric values for feature I1 and find the final, reduced number of intervals.
  - Reduce the number of numeric values for feature I2 and find the final, reduced number of intervals.
  - Reduce the number of numeric values for feature I3 and find the final, reduced number of intervals.

Computing GINI Index for Continuous Attributes

- 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

Cheat	No		No		No		Yes		Yes		Yes		No		No		No		No	
Taxable Income																				
Sorted Values →	60		70		75		85		90		95		100		120		125		220	
Split Positions →	55		65		72		80		87		92		97		110		122		172		230	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Yes	0	3	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0	3	0
No	0	7	1	6	2	5	3	4	3	4	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		*0.300*		0.343		0.375		0.400		0.420	