Knowledge Discovery and Data Mining

Unit # 3

Entropy-based Measure for Feature Ranking

- Similarity Measure
  - $S_{ij} = e^{-\alpha Dij}$
  - Where α = - (ln 0.5) / D
  - D is the average distance among samples in the data set
  - Normalized Euclidean distance measure is used to calculate the distance Dij between two samples xi and xj (n is the number of dimensions):
  - $Dij = \left[\sum_{k=1}^{n}((x_{ik} - x_{jk})/max_k - min_k))^2\right]^{1/2}$

Entropy-based Measure for Feature Ranking (Cont'd)

- Since all features are not numeric, the similarity for nominal variables is measured directly using Hamming distance.
- Sij = $\left(\sum_{k=1}^{n} |x_{ik} - x_{jk}| \right) / n$
- Where $|x_{ik} - x_{jk}|$ is 1 if $x_{ik} = x_{jk}$, and 0 otherwise.

- For mixed data, we can discretize numeric values and transform numeric features into nominal features before we apply this similarity measure.

Entropy-based Measure for Feature Ranking (Cont'd)

- The distribution of all similarities for a given data set is a characteristic of the organization and order of data in an n-dimensional space. This may be measured by entropy.
- The proposed technique compares the entropy measure for a given data set before and after removal of a feature. If the two measures are close, then the reduced set of features will satisfactorily approximate the original set.

- $E = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (S_{ij} \times \log S_{ij} + (1 - S_{ij}) \times \log (1 - S_{ij}))$

Sample	F1	F2	F3
R1	A	X	1
R2	B	Y	2
R3	C	Y	2
R4	B	X	1
R5	C	Z	3

	R1	R2	R3	R4	R5
R1		0/3	0/3	2/3	0/3
R2			2/3	1/3	0/3
R3				0/3	1/3
R4					0/3

Algorithm: Entropy based Ranking (Sequential Backward Ranking)

1. Start with the initial full set of features F.
2. For each feature f $\in$ F, remove one feature F and obtain a subset F_f. Find the difference between entropy for F and entropy for all F_f.
3. Let f_k be a feature such that the difference between entropy for F and entropy for f_{fk} is minimum.
4. Update the set of features F = F – {f_k}.
5. Repeat steps 2-4 until there is only one feature.

Entropy-based Feature Ranking Exercise

- Given four-dimensional samples where the first two dimensions are numeric and last two are categorical

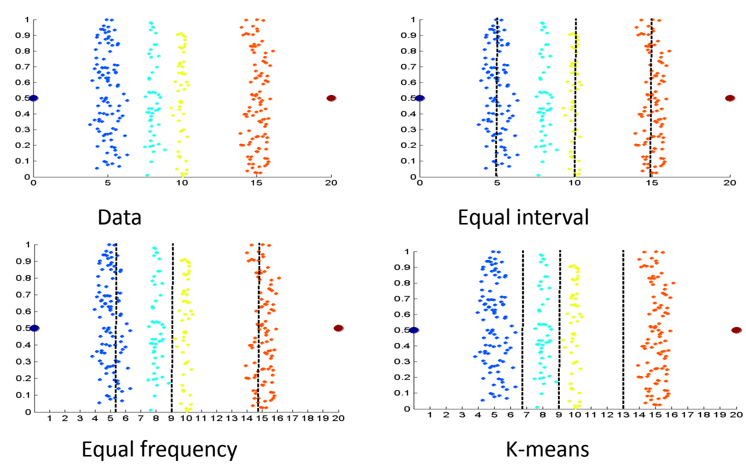
X1	X2	X3	X4
2.7	3.4	1	A
3.1	6.2	2	A
4.5	2.8	1	B
5.3	5.8	2	B
6.6	3.1	1	A
5.0	4.1	2	B

- Apply a method for unsupervised feature selection based on entropy measure to reduce one dimension from the given data set

Feature Discetization

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

Discretization Without Using Class Labels



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}
- If smoothing is performed
  - Mode: {1, 1, 1, 3, 3, 3, 5, 5, 5, 5}
  - Mean: {1.33, 1.33, 1.333 3, 3, 3, 5.25, 5.25, 5.25, 5.25}
  - Closest of the boundary value: {1, 1, 2, 3, 3, 3, 4, 4, 4, 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 I3 (in slide # 30, Unit # 2) using mean values as representatives for two bins.
  - The feature X2 (in slide # 6, Unit # 3) using closest boundaries for two bin representatives

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.