Knowledge Discovery and Data Mining

Unit # 7

Discretization using Weka

Demo

Entropy-based Measure for Feature Ranking

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

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

- $Sij = \left(\sum_{k=1}^{n} |x_{ik} - x_{jk}| \right) / n$
- 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.

Example

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

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