CS565: Data Mining

Fall 2007

Written Assignment 3

Due Date: December 5, 2007 in class.

Problem 1 (Based on exercise 7.16)

Many clustering algorithms can be applied to either numerical or categorical data but not on both of them. However, the EM algorithm can be easily extended to handle data that contain both numerical and categorical attributes. Explain how this is possible and show how to modify EM to achieve that.

Problem 2 (Based on exercise 8.7)

The idea of micro-clustering has been used to cluster data streams because it allows efficient on-line maintenance of clustering information. Using this idea, design an effective density-based clustering method for clustering evolving data streams.

Problem 3

Consider the following dataset of 6 records. Each record consists of 5 (binary) categorical attributes:

A	(1	0	1	1	0)
B	(1	1	0	1	1)
C	(1	0	1	1	0)
D	(0	1	0	1	0)
E	(1	0	1	0	1)
F	(0	1	1	1	0)

Also, given the following contingency table between two records R, S:

		R	
		1	0
S	1	a	b
	0	c	d

We define the following similarity measures:

$$SMC(R,S) = \frac{a+d}{a+b+c+d}$$

$$JC(R,S) = \frac{a}{a+b+c}$$

$$RC(R,S) = \frac{a}{a+b+c+d}$$

Using a hierarchical clustering algorithm produce the dendrograms for the following cases:

(a) Use the Single Link method (min distance/max similarity) with the SMC
(b) Use the Complete Link method (max distance/min similarity) with JC
(c) Use the Average Link method (average distance/average similarity) with RC