CMPT-740 Fall 2003
Foundations of Data Mining
Martin Ester

Assignment 3

Total marks: 60
Due date: October 8, 2003

Assignment 3.1

We want to cluster categoric data, i.e. data that have categoric attribute domains. The *k*-medoid algorithm can be applied to any datasets with a given pairwise distance function and, therefore, is applicable also to categoric data. The *k*-means algorithm, on the other hand, is much more efficient than the *k*-medoid algorithm, but it requires numeric data. The task of this assignment is to develop an analogon to the *k*-means algorithm for categoric data. We assume the following distance function for pairs of categoric objects:

$$dist(x,y) = \sum_{i=1}^{d} \delta(x_i, y_i) \;\; with \;\; \delta(x_i, y_i) = \begin{cases} 0 \;\; if \;\; x_i = y_i \\ 1 \;\; else \end{cases}$$

(a) What is the analogon for the means of a cluster *C* for catagoric data, i.e. a categoric object *m* minimizing the cluster cost

$$(*) \;\; TD(C,m) = \sum_{p \in C} dist(p,m)$$

Note that *m* must be computable by scanning the set of objects of *C* once (similar to the computation of the cluster means).

(b) Is the cluster representative *m* as defined in (a) unique?

(c) Describe the whole resulting algorithm for clustering categoric data, in particular the differences to the *k*-means algorithm.

Assignment 3.2

We analyse the density-based clustering algorithm DBSCAN.

(a) For which objects of a dataset does the cluster membership as determined by DBSCAN depend on the order of scanning the dataset? Give an example for two-dimensional data.

(b) How should we treat such objects in order to uniquely determine a DBSCAN clustering?

(c) The result of our refined DBSCAN algorithm does not depend on the order of scanning the dataset. Prove this proposition. Hint: prove the property of density-based clusters stated on slide 152 and use it as the basis for your argumentation.