ASSIGNMENT 3

DUE DATE: Wednesday 17 October, 2012 11:59 PM

In this assignment you will implement Fung et al.’s top-down specialization algorithm in R, and apply it to the given data set. The algorithm was discussed in class; however, it is advised that you also refer to the original paper for clarifications.


http://dl.acm.org/citation.cfm?id=1054047

The assignment will contribute 15 points to your final grade, but will be graded out of 75 points.

Algorithm

A brief pseudo code of the algorithm is given below. Nonetheless, you should refer to the paper and the lecture slides for details.

1. Let \( C \) be the cut containing the root nodes of the taxonomy trees
2. While there is a valid* and beneficial* node \( c \) in \( C \)
   a. Find the node \( c_{\text{best}} \) with the highest Score*
   b. In \( C \), replace \( c_{\text{best}} \) with its child nodes from the taxonomy tree
3. Output \( C \) and the data set after being generalized according to \( C \)

Note that the paper mentions efficient data structures to implement the algorithm (section 5.2 onwards). However, it will be sufficient if you simply use R operations to perform the steps of the algorithm without worrying about efficiency.

* see slides and paper for definition

Data set

The data set (adultcensus.data) that you will use to test your program is included in the ZIP file provided with this assignment (available from the assignment page of the course website). It is a downscaled version of the Census Income data set available from the UCI Machine Learning repository (http://archive.ics.uci.edu/ml/datasets/Adult).

The provided data set has the following quasi-identifiers: age, workclass, education, marital-status, occupation, race, sex, and native-country. In addition, there is a ninth attribute (salary-class) that should be used as the class label. The data is provided as a comma-separated table and can be easily read into R using the read.table() function.

Taxonomy tree

The taxonomy trees for the discrete attributes are provided in the included ZIP file. Note that the algorithm requires a taxonomy tree for each quasi-identifier (to generate a cut and then specialize attributes). However, no tree is specified for the age attribute in the included file. You may specify your own tree for this attribute, or use the method suggested by Fung et al. to handle continuous attributes (see Section 4).

Output

Given a value \( k \) for the k-anonymity requirement, your program should output (in two files):
1. the solution cut: how should each quasi-identifier be generalized, and
2. the generalized version of the given data set.

Tips

It would help you if you write functions to accomplish the following tasks:
• given a cut, obtain the generalized version of the data set,
• given a data set, obtain the size of the smallest equivalence class (see the code here: http://rwiki.sciviews.org/doku.php?id=tips:data:frames:count_and_extract_unique_rows),
• given a data set, an attribute $A$ and a value $a$, obtain the entropy in the class labels corresponding to the records in the data set that has $A=a$,
• given a data set, an attribute $A$ and a value $a$, determine if more than one class label is involved in the records in the data set that has $A=a$.

Grading and submission

The 75 points will be distributed as follows: correct implementation – 40; source code documentation – 15; clean R code – 20. Submit your R script file in blackboard by the stated deadline.

You must work alone in this assignment. It is okay to discuss the algorithm and the paper with your peers, but the R scripts must be written individually. Start early!