Programming Exercise **T-122.102, Spring 2003**

Select one of the tasks below. Write a report of about 3–4 A4 pages (plus any appendices such as program listings, large tables of results, etc). The emphasis should be more on the discussion in your report than on the programming work. In other words, don't stop once you have a working program and have run it on some data — then you should spend time thinking about your results, and perhaps doing more experiments. You are encouraged to use any implementations of algorithms that can be found on the web, but remember to cite all sources. Turn in your report to the course assistant by May 30. Acceptable delivery formats are, in decreasing order of preference, (i) physically on paper to Jouni Seppänen, PL 5400, 02015 TKK; (ii) as a PDF or PostScript file to `Jouni.Seppanen@hut.fi`.

1. Compare at least three different frequent itemset mining algorithms. Use artificial data sets; try to find for each algorithm a data set on which that algorithm is better than the other algorithms. Some implementations by Bart Goethals can be found on his web page at `http://www.cs.helsinki.fi/u/goethals/software/`. The paper at `http://citeseer.nj.nec.com/zheng01real.html` may be of interest.

2. Investigate mining frequent itemsets in different clusters of a data set. Ask the course organizers for a copy of the paper

   > Hollmén, Seppänen, Mannila. Mixture models and frequent sets: combining global and local methods for 0–1 data. To appear in SIAM Data Mining 2003.

   and base your work on it. (Feel free to criticize the paper!) You can use e.g. Goethals's Apriori implementation (see task 1) and Carreira-Perpiñán's Bernoulli mixtures implementation at `http://cns.georgetown.edu/~miguel/software.html`. Use your own data, or see e.g. the various data sets linked from Jon Kleinberg's page at `http://www.cs.cornell.edu/Courses/cs685/2002fa/`. For example, the "bibliography" data set could be transformed into binary form using the bag-of-words approach.

3. Implement the algorithm described in

   > Ian T. Nabney. Efficient training of RBF networks for classification. ICANN'99, pp. 210–215. `http://www.ncrg.aston.ac.uk/cgi-bin/tr_avail.pl?trnumber=NCRG/99/002`

   and try it on various classification data sets from the UCI repository at `http://www.ics.uci.edu/~mlearn/MLRepository.html`. Implement and test also some other model using the same way of transforming a linear output to binary.

4. Select a topic of your own in the scope of the course. Before investing a lot of effort on the topic, write an abstract of one or two paragraphs and show it to the course organizers.