Introduction to Data Mining - Fall 2009

1 Projects

1. *Classification for rare-class problems* A comparative study of different classification techniques to analyze rare class problems. You are to prepare a survey of the enclosed literature and give a presentation in class.

   - Joshi, M.V., and Agrawal, R., PNrule: A New Framework for Learning Classifier Models in Data Mining (A Case-study in Network Intrusion Detection) ¡http://citeseer.ist.psu.edu/agarwal00pnrule.ht (2001)
   - Joshi, M.V., Agrawal, R., and Kumar, V., Mining Needles in a Haystack: Classifying Rare Classes via Two-Phase Rule Induction ¡http://www-users.cs.umn.edu/
   - Joshi, M.V., Kumar, V., Agrawal, R., Predicting Rare Classes: Can Boosting Make Any Weak Learner Strong? ¡http://www-users.cs.umn.edu/
   - Joshi, M.V., Kumar, V., Agrawal, R., On Evaluating Performance of Classifiers for Rare Classes (2002) ¡http://www-users.cs.umn.edu/

2. *Scalable clustering algorithms* A comparative study of scalable data mining techniques. Read the papers below and prepare the talk and report of main results discussed in the papers.

   - Tian Zhang, BIRCH: An Efficient Data Clustering Method for Very Large Databases - ¡http://citeseer.ist.psu.edu/zhang99birch.html¿. 1999
   - Ganti, Ramakrishnan, Clustering Large Datasets in Arbitrary Metric Spaces ¡http://citeseer.ist.psu.edu/ga 1998
   - Bradley, Fayyad, Reina Scaling Clustering Algorithms to Large Databases ¡http://citeseer.ist.psu.edu/bradl 1998
   - Farnstrom, Lewis, Elkan, Scalability for Clustering Algorithms Revisited ¡http://citeseer.ist.psu.edu/farnstr ¡http://citeseer.ist.psu.edu/farnstrom00scalability.html¿- 2000 ¡http://citeseer.ist.psu.edu/farnstrom00scal

3. *Comparison of effectiveness of three classification algorithms: Decision tree and Naive Bayesian.* You will be given a set of data. You first discretize continues attributes. After that you randomly select the 2/3ds of your data and design a decision tree. After that you classify the remaining data and compare the number of classification errors of each of the algorithms. The data you will be given there are attributes that are not defined. You should ignore these attributes.

4. *Implementation of A'priori algorithm for Associative rules mining* You are to implement A'Priori algorithm. You wo'ill be given a set of data and two numbers: support level and confidence level. You should generate the set of rules for your data. Before you do that you should eliminate highly correlated data and assume the rest of the attributes are either independent or have very low correlation.