# Projects for CSC 422 & 522

Students are encouraged to collaborate with up to two other students on a data mining project. Each group must submit a written project proposal partway through the semester (**due date September 20**), and a project report and a poster (**due date for both November 29**). At the end of the semester we will have a poster presentation and award a prize for the best poster. Students enrolled in 522 must (additionally!) submit a full research paper (8-15 pages) describing their project (**due date November 29**). Every student must submit his/her own paper, teamwork is only allowed for the research project, not for the paper write-up.

**Data sets:** Here is a list of web pages with links to datasets that you can choose for your project to work on. **You can also propose data of your own choice.**

http://www.kdnuggets.com/datasets/

http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html

http://archive.ics.uci.edu/ml/

**Project ideas** together with related literature: You are encouraged to choose your own research project, but just in case, here are some ideas.

1. Bioinformatics data analysis
   Identify and model biases in next-generation sequencing.
   Data source: TBA
   References:
   Nucleic Acids Res. 2010 Jul 1;38(12):e131. Epub 2010 Apr 14.
   Biases in Illumina transcriptome sequencing caused by random hexamer priming.
   Hansen KD, Brenner SE, Dudoit S.
   Nucleic Acids Res. 2008 Sep;36(16):e105. Epub 2008 Jul 26.
   Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.
   Dohm JC, Lottaz C, Borodina T, Himmelbauer H.

2. Evaluating Performance of Classifiers
   Compare the bias and variance of models generated using different evaluation methods (leave one out, cross validation, bootstrap, stratification, etc.)
   References:
   a. Kohavi, R., A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection (1995)

b Efron, B. and Tibshirani, R., Cross-Validation and the Bootstrap: Estimating the Error Rate of a Prediction Rule (1995)
c. Martin, J.K., and Hirschberg, D.S., Small Sample Statistics for Classification Error Rates I: Error Rate Measurements (1996)
d. Dietterich, T.G., Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms (1998)

2. Support Vector Machine (SVM)
   Present an overview of SVM or applying Support Vector Machines to various application domains.
   References:
   a. Mangasarian, O.L., Data Mining via Support Vector Machines (2001)
   b. Burges, C.J.C., A Tutorial on Support Vector Machines for Pattern Recognition (1998)
   c. Joachims, T., Text Categorization with Support Vector Machines: Learning with Many Relevant Features (1998)
   d. Salomon, J., Support Vector Machines for Phoneme Classification (2001)

3. Cost-sensitive learning
   A comparative study and implementation of different techniques for ensemble learning such as bagging, boosting, etc.
   References:
   a. Freund Y. and Schapire, R.E., A short introduction to boosting (1999)
   b. Joshi, M.V., Kumar, V., Agrawal, R., Predicting Rare Classes: Can Boosting Make Any Weak Learner Strong? (2002)
   c. Quinlan, J.R., Boosting, Bagging and C4.5 (1996)
   d. Bauer, E., Kohavi, R., An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants (1999)

4. Semi-supervised learning (classification with labeled and unlabeled data)
   Applying different semi-supervised learning techniques to UCI data sets.
   References:
   a. Nigam, K., Using Unlabeled Data to Improve Text Classification (2001)
   b. Seeger, M., Learning with labeled and unlabeled data (2001)
   c. Nigam, K. and Ghani, R., Analyzing the Effectiveness and Applicability of Co-training (2000)
   d. Vittaut, J.N., Amini, M-R., Gallinari, P., Learning Classification with Both Labeled and Unlabeled Data (2002).

5. Classification for rare-class problems
   A comparative study and/or implementation of different classification techniques to analyze rare class problems
   References:

a. Joshi, M.V., and Agrawal, R., PNrule: A New Framework for Learning Classifier Models in Data Mining (A Case-study in Network Intrusion Detection) (2001)
b. Joshi, M.V., Agrawal, R., and Kumar, V., Mining Needles in a Haystack: Classifying Rare Classes via Two-Phase Rule Induction (2001)
c. Joshi, M.V., Kumar, V., Agrawal, R., Predicting Rare Classes: Can Boosting Make Any Weak Learner Strong? (2002)
d. Joshi, M.V., Kumar, V., Agrawal, R., On Evaluating Performance of Classifiers for Rare Classes (2002) (2002)

6. Time Series Prediction/Classification
A comparative study and/or implementation of time series prediction/classification techniques
References:
a. Geurts, P., Pattern Extraction for Time Series Classification (2001)
b. Kadous, M.W., A General Architecture for Supervised Classification of Multivariate Time Series (1998)
c. Giles, C.L., Lawrence, S. and Tsoi, A.C., Noisy Time Series Prediction using a Recurrent Neural Network and Grammatical Inference (2001)
d. Keogh, E.J. and Pazzani, M.J., An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback (1998)
e. Chatfield, C., The Analysis of Time Series, Chapman & Hall (1989)

7. Sequence Prediction
A comparative study and implementation of sequence prediction techniques
References:
a. Laird, P.D., Saul, R. Discrete Sequence Prediction and Its Applications. Machine Learning, 15(1): 43-68 (1994)
b. Sun, R. and Lee Giles, C., Sequence Learning: From Recognition and Prediction to Sequential Decision Making (2001)
c. Lesh, N., Zaki, M.J., and Ogihara, M., Mining features for Sequence Classification (1999)

8. Association Rules for Classification
A comparative study and implementation of classification using association patterns (rules and itemsets)
References:
a. Liu, B., Hsu, W., and Ma, Y., Integrating Classification and Association Rule Mining (1998)
b. Liu, B., Ma, Y. and Wong, C-K, Classification Using Association Rules: Weaknesses and Enhancements (2001)
c. Li, W., Han, J. and Pei, J., CMAR: Accurate and Efficient Classification Based on Multiple Class-Association (2001)
d. Deshpande, M. and Karypis, G., Using Conjunction of Attribute Values for Classification (2002)

9. Spatial Association Rule Mining

   A comparative study on spatial association rule mining.

   References:

   a. Koperski, K., and Han, J., Discovery of Spatial Association Rules in Geographic Information Databases (1995)
   b. Shekhar, S. and Huang, Y., Discovering Spatial Co-location Patterns: A Summary of Results (2001)
   c. Malerba, D., Esposito, F. and Lisi, F., Mining Spatial Association Rules in Census Data (2001)

10. Temporal Association Rule Mining

   A comparative study and/or implementation of temporal association rule mining techniques

   References:

   a. Li, Y., Ning, P., Wang, and S., Jajodia, S., Discovering Calendar-based Temporal Association Rules (2001)
   b. Chen, X. and Petrounias, Mining temporal features in association rules
   c. Lee, C.H., Lin, C.R. and Chen, M.S., On Mining General Temporal Association Rules in a Publication Database (2001)
   d. Ozden, B., Ramaswamy, Silberschatz, Cyclic Association Rules (1998)
   e. Literature on Sequential Association Rule Mining below

11. Sequential Association Rule Mining

   A comparative study and/or implementation of sequential association rule mining techniques

   References:

   a. Srikant, R. and Agrawal, R., Mining Sequential Patterns: Generalizations and Performance Improvements (1996)
   b. Mannila, H. and Toivonen, H., Verkamo, A.I., Discovery of Frequent Episodes in Event Sequences (1997)
   c. Joshi, M., Karypis, G., and Kumar, V., A Universal Formulation of Sequential Patterns (1999)
   d. Borges J., and Levene, M., Mining Association Rules in Hypertext Databases (1998)

12. Outlier Detection

   A comparative study and/or implementation of outlier detection techniques.

   References:

   a. Knorr, Ng, A Unified Notion of Outliers: Properties and Computation, - 1997
   b. Knorr, Ng, Algorithms for Mining Distance-Based Outliers in Large Datasets - 1998
   c. Breunig, Kriegel, Ng, Sander, LOF: Identifying Density-Based Local Outliers - 2000
   d. Aggarwal, Yu, Outlier Detection for High Dimensional Data – 2001
   e. Tang, Chen, Fu, Cheung, A Robust Outlier Detection Scheme for Large Data Sets - 2001

15. Scalable clustering algorithms
    A comparative study of scalable data mining techniques.
    References:
    a. Tian Zhang, BIRCH: An Efficient Data Clustering Method for Very Large Databases - . 1999
    b. Ganti, Ramakrishnan, Clustering Large Datasets in Arbitrary Metric Spaces, 1998
    c. Bradley, Fayyad, Reina Scaling Clustering Algorithms to Large Databases –1998
    d. Farnstrom, Lewis, Elkan, Scalability for Clustering Algorithms Revisited - 2000
16. Clustering association rules and frequent item sets
    A comparative study of techniques for clustering association rules.
    References:
    a. Toivonen, Klemettinen, Pruning and Grouping Discovered Association Rules, 1995

    b. Widom, Clustering Association Rules - Lent, Swami - 1997
    c. Gunjan K. Gupta , Alexander Strehl AND Joydeep Ghosh, Distance Based Clustering of Association Rules

**Project proposal.** The project proposal has two primary purposes. First, it ensures each student has an identified a concrete area of interest early on in the course. This aids understanding of course material by providing a specific problem of interest, and provides time for data acquisition. Second, the proposal provides the instructor with a basis for offering advice, especially in cases in which he judges the proposed analysis either trivial or infeasible, or in which the proposed analysis involves confidential or proprietary information. Accordingly, the instructor will not read "drafts" of proposals; he will be happy to read or discuss revised conceptions of projects, but will only grade the first version given to him.

The primary factor in the grading of proposals is whether the proposal provides the information needed by the instructor to make these judgments and offer this advice. The instructor expects to recommend changing (narrowing or broadening) the proposed analysis in some cases, and this normally does not affect the grade on the proposal. What does affect the grade is a proposal that does not clearly provide the information needed to judge the proposed work.

Toward these ends, the proposal should briefly describe

- What application area the project addresses;
- What type of knowledge is sought, and why (its intended use and benefit);
- What data sources are available; and
- What data sources are expected to be used, and why.

The proposal may additionally describe any expectations about details of the analysis plan and methods to be applied; such expectations do not constitute commitments, but can improve the proposal quality if done well. In the best cases, the project proposal constitutes the initial portion of the project report.

There is no set length for project proposals, but 2-3 pages should be adequate, or a bit longer if data dictionary tables are included. Mere length does not make for better proposals, unless for some reason the proposal cannot state the indicated information succinctly. Excessive length suggests failure to grasp the idea of a “proposal”; please consider whether all the material provided is really appropriate for a proposal rather than for the eventual report. Here a short course about how to write a proposal http://foundationcenter.org/getstarted/tutorials/shortcourse/index.html

**Result report: TBA**

**Poster: TBA**

**Students enrolled in 522** must (additionally!) submit a full research paper (8-15 pages) describing their project. **Every student must submit his/her own paper, teamwork is only allowed for the research project, not for the paper write-up.** Here some guidelines about

- How to write a paper? The Scientific Paper
- How NOT to write a paper
- How to write a paper in Scientific Journal Style and Format
- SCIgen - An Automatic CS Paper Generator