

Guidelines and Deadlines for Data Mining Term Project

1. Guidelines

You are encouraged to work in groups of two to four for the term project. This project is a critical part of the course, and a significant factor in determining your grade. First, each group should submit (hardcopy) a one/two page proposal summarizing the proposed project including the plan of attack and at least two key references, on or before **March 25**, and then give a 3-4 minute presentation of your plan in class on April 1st. Please feel free to discuss your project with me before that. Also, if you have difficulty in finding a project partner, you could contact me with a list of your interest area(s), so I can try to match you up with another student before the March 27th deadline. Each group shall give a 15–20 minute presentation on their project around mid-late April

The **deadline** for the hardcopy term paper submission is **May 7th**. This copy should be 10-30 pages (1.5 spacing) including figures, tables and/or references. If a part of this is an html file/directory, you could just give me a pointer to the URL, and reduce your printed submission accordingly. One submission per group.

In some cases I may provide feedback on the final term paper by May 9th, and give you the chance to submit an improved version by May 12th.

2. Project Types

The project can be a practical one (a specific application of data mining), or a theoretical one (propose a new algorithm etc). In-depth survey papers on new or focussed topics are also possible, but it should be an original and unique niche (several broad surveys exist on the web; a cut-and-paste job will not do!).

You can choose any topic relating to data mining, and use any reasonably large data set. Some tools and datasets are referred to at the course website; also see kdnuggets.com.

Project Type A: (Public Domain Data Based). There are many public domain datasets. Several of these are part of competitions such as the KDD cup, and for some can also find papers on how others have fared on these data sets. Note that some of these data sets are very difficult, but still worth exploring. (for some current competitions such as Netflix there is also prize money!). We have compiled a list of some of these at <http://www.ideal.ece.utexas.edu/~ghosh/DM-Competition.html>

A rather fascinating compilation of datasets is at:

<http://www.datawrangling.com/some-datasets-available-on-the-web.html>

Project Type B (Algorithm Driven).

Below are some (other) suggested topics; by using Google and **Citeseer** <http://citeseer.ist.psu.edu/cs>, you should be able to find some materials on them.

Clustering: - clustering with constraints; clustering very high dimensional data; clustering of gene sequences; etc, see <http://ideal.ece.utexas.edu/~ghosh/ch8s.ps> for a review and other issues.
- clustering tools: (a) CLUTO (Karypis): understand this tool; add to its functionality, or
(b) Co-clustering: simultaneous clustering of rows and columns is a technique with rich possibilities; you can add to its functionality and do empirical studies. Or look at bioinformatics applications. (code at <http://www.cs.utexas.edu/dml/Software/cocluster.html>)

Classification: - comparing different approaches to solving multi-class problems: error correcting output codes (Bakiri & Dietterich) vs. Binary Hierarchical Classifier (Kumar, Ghosh, Crawford).
- bagging with strong learners vs. boosting with weak learners
- dealing with highly different costs; priors (Charles Elkan, Domingos)

Mining Structured Data: where data has some special structure, e.g. graphs, XML documents, etc. (See <http://hms.liacs.nl/index.html>). A related topic (where the i.i.d. assumption is again violated) is:

Statistical Relational Learning: Methods of directly working on relational databases, e.g. Markov Logic Networks (MLN) and PRMs (Koller). The MLN folks have a public-domain software, *Alchemy* <http://www.cs.washington.edu/ai/alchemy/> that you can play with. (also see video tutorials at: http://videolectures.net/icml05_getoor_srl1/; http://videolectures.net/icml07_domingos_psr/)

Mining Software Repositories: for maintenance, improve software design etc. See proceedings or a workshop on this topic at: msr.uwaterloo.ca Some data for GCC, cleaned up for Weka, is available. The KDnuggets site has a link to repositories of SourceForge docs.

Bioinformatics: Hot topic! e.g. see the KDD2002 challenge competition <http://www.biostat.wisc.edu/~craven/kddcup/>.

Streaming Data; Change Detection Analyzing data that you see only once (Johannes Gehrke); Detecting and modeling "change" in time sequence data, e.g. those gathered from networked computer systems; change in customer segmentation (how do clusters move, get created/die, as the stats of data gathered changes over time?). I am particularly interested in ensemble approaches to on-line learning. A nice overview is given in: Kuncheva L.I. Classifier ensembles for changing environments, Proc. 5th International Workshop on Multiple Classifier Systems, MCS2004, Cagliari, Italy, in F. Roli, J. Kittler and T. Windeatt (Eds.), Lecture Notes in Computer Science, Vol 3077, 2004, 1-15. PDF of talk copied at: <http://www.ideal.ece.utexas.edu/~ghosh/mcs04Kuncheva.pdf>

Privacy Perserving Data Mining: e.g. see paper No.2 at <http://www.ideal.ece.utexas.edu/~srujana/papers.html>, and the special issue of SIGKDD: <http://www.acm.org/sigkdd/explorations/issue4-2.htm>

Or you can consider Distributed Data Mining: How to do data mining (regression, classification, AR, clustering) over multiple, geographically distributed and possibly quite varying, data sets.

Remote Sensing: (i) predicting forest cover type from satellite images (see UCI).
(ii) classifying land cover from hyper-spectral data (I can provide).
(iii) discovering and modelling heterogenous regions in spatial data
(e.g. <http://www.cs.umn.edu/research/shashi-group/> ;
<http://www.dbs.informatik.uni-muenchen.de/Forschung/KDD/SpatialKDD/>) more generally modeling signals with distinct regimes (determine boundaries; model each segment).

Direct Marketing/Market Research: e.g. donor database (KDD cup 1998), or working with the ERIM dataset <http://research.chicagogsb.edu/marketing/databases/erim/index.aspx>

Web Mining: There is a fairly extensive list of projects provided on web mining (along with references, etc) at <http://ideal.ece.utexas.edu/course/prac/03sp/prac-projects.html> Also the KDD Cup for 2003 related to citation prediction, see <http://www.cs.cornell.edu/projects/kddcup/>

Statistical Angles Explore role of sampling techniques; make connections between statistical and machine learning approaches, e.g. see Trevor Hastie's works.

Semiconductor Manufacturing: measuring quality of chips, equipment, process etc. from process logs. (you need to get your own data).

Augmenting WEKA: with new functionality.