DATA MINING LAB

B.Tech. IV Year I Sem.   **L   T   P   C**
Course Code: CS703PC   **0   0   3   2**

Course Objectives:
- To obtain practical experience using data mining techniques on real world data sets.
- Emphasize hands-on experience working with all real data sets.

List of Sample Problems: Task 1: Credit Risk Assessment

Description:
The business of banks is making loans. Assessing the credit worthiness of an applicant is of crucial importance. You have to develop a system to help a loan officer decide whether the credit of a customer is good, or bad. A bank's business rules regarding loans must consider two opposing factors. On the one hand, a bank wants to make as many loans as possible. Interest on these loans is the banks profit source. On the other hand, a bank cannot afford to make too many bad loans. Too many bad loans could lead to the collapse of the bank. The bank's loan policy must involve a compromise: not too strict, and not too lenient.
To do the assignment, you first and foremost need some knowledge about the world of credit. You can acquire such knowledge in a number of ways.
1. Knowledge Engineering. Find a loan officer who is willing to talk. Interview her and try to represent her knowledge in the form of production rules.
2. Books. Find some training manuals for loan officers or perhaps a suitable textbook on finance. Translate this knowledge from text form to production rule form.
3. Common sense. Imagine yourself as a loan officer and make up reasonable rules which can be used to judge the credit worthiness of a loan applicant.
4. Case histories. Find records of actual cases where competent loan officers correctly judged when, and when not to, approve a loan application.

The German Credit Data:
Actual historical credit data is not always easy to come by because of confidentiality rules. Here is one such dataset, consisting of 1000 actual cases collected in Germany. Credit dataset (original) Excel spreadsheet version of the German credit data.
In spite of the fact that the data is German, you should probably make use of it for this assignment. (Unless you really can consult a real loan officer!)
A few notes on the German dataset
1. DM stands for Deutsche Mark, the unit of currency, worth about 90 cents Canadian (but looks and acts like a quarter).
2. owns_telephone. German phone rates are much higher than in Canada so fewer people own telephones.
3. foreign_worker. There are millions of these in Germany (many from Turkey). It is very hard to get German citizenship if you were not born of German parents.
4. There are 20 attributes used in judging a loan applicant. The goal is to classify the applicant into one of two categories, good or bad.