Department of Computer Science and Engineering
University of California at San Diego

CSE 255
Winter 1998

Project 2

DUE AT THE START OF CLASS, TUESDAY FEBRUARY 3.

The assignment here is to go through the stages of a data mining project, using the MineSet software from Silicon Graphics and data acquired through the worldwide web. The deliverable for this assignment is a report describing your project. The report should discuss the goals of the project, the data that you collected, the tools that you applied to the data, the knowledge that was acquired from the data, how this knowledge was validated, and how the knowledge could be used.

For an example of an applied AI project where two data mining methods and hand-coded knowledge were compared, see *Expert prediction, symbolic learning, and neural networks: An experiment on greyhound racing* by Hsinchun Chen *et al.*, *IEEE Expert*, December 1994, **9**(6)21–27. This paper is available at `http://ai.bpa.arizona.edu/papers/`.

It is important to have clear goals for your project before undertaking it. Whether you use supervised learning or unsupervised learning, you should have a vision for how the acquired knowledge can be applied for a commercial or industrial purpose. This vision should be explained with precision and concreteness in the report.

Validation of the learned knowledge is another important issue. For example, if you discover patterns in a historical database, how can you be sure that the patterns will recur in future data? You must also think carefully about data quality and data "cleaning" and preprocessing. How will you deal with missing data? Will missing data bias your conclusions? How will you detect and/or deal with erroneous data?

As in the real world, your choice of application domain will depend heavily on the availability of usable data. Combining existing datasets is often key to obtaining novel results. Many interesting datasets are available in a U. C. Irvine repository linked to the class web page. However these datasets are mostly small, and not from commercial or industrial sources.

As before, you should document your work in a report of eight to ten pages. The report should discuss the results of each of the stages of a data mining project as outlined in class. The quality of writing in the report must be good: be sure to follow carefully the guidelines in the references on technical writing linked to the class web page.

Assistance

For this project help with the MineSet software is available from Dave Martinez, an M.S. graduate of the Program in Advanced Manufacturing. Dave can be contacted at `damartin@sdsc.edu` and (619) 534-8395.

MineSet only runs on the Silicon Graphics workstation named `mina` in the INC laboratory on the 6th floor of the APM building. The permanent researchers in this room have priority access to resources and the right not to be disturbed. Please be courteous and do all possible work away from the lab. Do not hold team meetings in the lab, and do not have loud or lengthy discussions there. Plan your work carefully to minimize the amount of time that you need to use `mina`. Do all web searching, programming, and other general exploration elsewhere.