Data Mining and Knowledge Discovery

Department of Information Management, NSYSU

Spring, 2007 Syllabus

Goal

- Present fundamental concepts and techniques for data mining
- Provide necessary background for applying data mining to business problems
- Conduct case studies on real data mining examples
- Practice data mining tools on real data

Description

This course will focus more on the technical aspects of data mining rather than its business application, though a certain case studies will be conducted in the class and by the students. The lecture will be delivered in English. However, students can raise their questions in either Chinese or English, and the oral presentation can be conducted in either Chinese or English. Previous lecture files (presented in Chinese) are available at http://140.117.74.227/ under the directory of Data Mining.

Instructor/TA

- Prof. San-Yih Hwang (黃三益)
  office: 管 4071
  extension: 4723
  email: syhwang@mis.nsysu.edu.tw
- Teacher assistant to be announced

Class meeting time

- 2:10-5pm, Monday

Class meeting Place

- 管 3019

Textbook

Tan, Steinbach, Kumar, *Introduction to Data Mining*, Addison Wesley, 2006, 歐亞代理 (http://www-users.cs.umn.edu/~kumar/dmbook/) , slides available at http://www-users.cs.umn.edu/~kumar/dmbook/index.php#item4

Referenced books

- D.J. Hand, H. Mannila, and P. Smyth, *Principle of Data Mining*, MIT Press, 2001. (good on the theoretical aspects of data mining/knowledge discovery)
- M. J. A. Berry and G. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, Wiley Computer Publishing, 2nd. ed., 2004. (good on the business applications of data mining/knowledge discovery)

Prerequisites

You must have background in data structure and basic statistics, and some programming experiences.

Homeworks NONE. But at the end of each class, 2 or 3 exercises about the topic covered in next class will be given. You are required to preview at least half of the materials and do the exercises before each class.

Quiz

- There will be four quizzes. The purpose of each quiz is to evaluate your understanding on how various data mining techniques work. Before each quiz, a set of exercises will be given (without having to turn them in) and the quiz will be VERY similar to one of the given exercises.

Projects

- There will be TWO projects: mid-term project and final project. The purpose of projects is to let you get familiar with data mining methodologies and data mining tools. The mid-term project asks you to study a real data mining case, which will be handed out in a few weeks. The outcome of this project includes a written report (in English) and an oral presentation in class (in either Chinese or English). The final project asks you to conduct a data mining task based on some real data that is accessible to you. You should apply some data mining tool, either a free tool listed at the appendix or a commercial tool that is accessible to you, on the data set. The deliverables of this project include a written report (in English) and an oral presentation in class (in either Chinese or English). Both projects are conducted on a group basis. Each group should have no more than three members.

Exams NONE

Grading Quiz 12% each, mid-term project: 20%, final Project 25%, Class

Discussion: 7%.

Class Schedule

Date	Topics	Events
2/26	Chapter 1 Introduction	
3/5	Chapter 2 Data	
3/12	Chapter 3 Exploring data, Chapter 4 Classification	Project 1 description handed out
3/19	Chapter 4 Classification	
3/26	Chapter 4 Classification	Quiz 1
4/2	OFF	
4/9	Data Mining Methodologies ((supplement)	
4/16	Chapter 6 Association analysis	
4/23	Chapter 8 Clustering	
4/30	Mid term project presentation	Project 1 due Project 2 description handed out
5/7	Chapter 6 Clustering	Quiz 2
5/14	Chapter 5 Advanced classification	
5/21	Chapter 7 Advanced association	
5/28	Chapter 9 Advanced clustering	
6/4	Chapter 10 Anomaly detection	Quiz 3
6/11	Recommendation (supplement)	
6/18	Final project presentation	Project 2 due

RESOURCES

A. SOFTWARE

General Purpose Data Mining	• WEKA (Source: Java) • RapidMiner • MLC++ (Source: C++) • SIPINA • List from KDNuggets (Various) • List from Data Management Center (Various)
Classification	• C4.5 (Decision tree) • OC1 (Oblique decision tree) • Ripper (Rule-based) • CBA (association-rule based) • bayes (Naive Bayes) • Evidential distance-based (nearest-neighbor) • PEBLS (nearest-neighbor) • mlp (Neural Network) • tiberius (Neural Network) • svmlight (Support Vector Machine)
Association Analysis	• FIMI Repository of Algorithms • Apriori, Eclat, and FP Growth • ARTool • ARMADA (Association rule mining in Matlab) • Tree Mining, Closed Itemsets, Sequential Pattern Mining • Tree Mining, Closed Itemsets, Sequential Pattern Mining • PAFI
Cluster Analysis	• CLUTO • Open Source Clustering Software • Model-based Clustering • Online software for Clustering
Anomaly Detection	• ORCA (distance based)
Regression	• Regression routines

Data Preprocessing	• Feature Selection • Isomap (Dimensionality Reduction - in Matlab)

B. Data Sets

- IDS data sets
- Data Sets for Data Mining
- Competition Data Set
- UCI Machine learning repository
- Quest data repository
- KDNuggets