CSE634 Data Mining, Spring 2018
Professor Anita Wasilewska

web page: http://www.cs.stonybrook.edu/~ cse634/

Meets Monday, Friday 5:30 pm - 6:50

Place JAVITS 111

Professor Anita Wasilewska

e-mail address: anita@cs.stonybrook.edu

Office phone number: 632 8458

Office location: New Computer Science Department, Room 208

Professor Office Hours Monday, Wednesday 7:15 pm - 8: 30pm, and by appointment.

TA t.b.a

Textbook

DATA MINING Concepts and Techniques
Jiawei Han, Micheline Kamber
Morgan Kaufman Publishers, 2003
Second Edition

http://web.engr.illinois.edu/~ hanj/bk2/

Course Description

Data Mining (DM), called also Knowledge Discovery in Databases (KDD) and now called also BIG DATA is a new multidisciplinary field. It brings together research and ideas from database technology, machine learning, neural networks, statistics, pattern recognition, knowledge based systems, information retrieval, high-performance computing, and data visualization. Its main focus is the automated extraction of patterns representing knowledge implicitly stored in large databases, data warehouses, and other massive information repositories.

The course will closely follow the book and is designed to give a broad, yet in-depth overview of the Data Mining field and examine the most recognized techniques in a more rigorous detail. It also will explore the newest trends and developments of the field in form of students talks based on newest research papers from the field.

Course Structure

Part 1 Introduction; Data Preprocessing

Book chapters 1, 2 and Lectures 1, 2

Part 2 Classification

Decision Tree Induction and Neural Networks

Book chapter 6 and Lectures 3 - 7

Team Classification Project

See the Project Description section and check the link on the course Website.

Part 3 Association Analysis

Apriori Algorithm

Classification by Association

Book chapter 5 and Lectures 8, 9

Test Review One

Lecture 10

Part 4 Genetic Algorithms

Genetic Algorithms Introduction

Genetic Algorithms Examples

Book chapter 6, Lectures 11, 12

Test Review Two

Lecture 13

Midterm/Final Test

It is in class test and covers material from **Parts 1- 4**

Part 5 Cluster Analysis

Book chapter 7 and Lectures 14, 15

Part 6 Foundations of Data Mining

Lecture 16

Part 7 Students **Research Presentations**

Attention: Project and Research Presentations are to be conducted in **teams**

TEAMS consists of 3 students and the SAME for Classification Project and Presentations

Grading Components

During the semester students are responsible for the following (in order as listed).

1. Team Project (40pts)

2. Midterm/Final Test (70pts)

3. Team Research Presentation (60pts)

4. Final Report (30 points).

FINAL GRADE COPMUTATION

Attention: **NONE of the grades will be curved**

During the semester you can earn 200pts or more (in the case of extra points).

The % grade will be determine in the following way: # of earned points divided by 3 = % grade.

The % grade which is **translated** into letter grade as follows.

100 - 90 % is A range:

A (100-96%), A- (95- 90%),

89 - 80 % is B range:

B- (80 - 82%), B (83 -85%), B+ (86 -89%),

79 - 70 % is C range:

C- (70- 72%), C (73-75%), C+(76-79%),

69 - 60 % is D range, and

F is below 60%.

Course Contents

The course will follow the book very closely and in particular we will cover all or some of following chapters and subjects. The order does not need to be sequential.

Chapter 1 Introduction. General overview: what is Data Mining, which data, what kinds of patterns can be mined.

Chapter 2 Data preprocessing: data cleaning, data integration and transformation, data reduction, discretization and concept hierarchy generation.

Chapters 3, 4 Data Warehouse and OLAP technology for Data Mining. (Students presentations)

Chapter 5 Mining Association Rules in Large Databases. Transactional databases and Apriori Algorithm (LECTURE and Students Presentation).

Chapter 6 Classification and prediction.

1. Decision Tree Induction ID3, C4.5). (Lecture and Students Presentations)

2. Neural Networks (Lecture and students Presentations)

3. Bayesian Classification. (Lecture and Students presentations).

4. Classification based on Concepts from Association rule mining (Lecture)

5. Genetic algorithms. (Lecture and Students presentations)

6. Statistical Prediction (Students presentations).

Chapter 7 Cluster Analysis. A Categorization of major Clustering methods. (Lecture and Students presentations).

Applications and TRENDS in DM - chapters 8 -11, reading and /or students presentations.

Foundations of Data Mining

SPRINGER Encyclopedia of Complexity and Systems Science, 2009 Editors: Editor-in-chief: Meyers, Robert A http://www.springer.com/us/book/9780387758886

PROJECT DESCRIPTION

Project **goal** is to use Internet based Classification Tools to build two type classifiers: **descriptive** and **non-descriptive**. Discuss the results. **Compare** these two approaches on the basis of obtained results.

1. Descriptive Classifier

Use a **Decision Tree** tool to generate sets of **discriminant rules** describing the content of the data.

Use WEKA:
http://www.cs.waikato.ac.nz/˜ ml/weka/index.html)

2. Non-Decsriptive Classifier

Use **Neural Networks** tool to build your Classifier

Use WEKA or a tool of your choice. Describe specifics of your tool in a way that makes your report comprehensible for others.

Here are some tools suggestions:

http://www.mathworks.com/products/neural-network/?requestedDomain=www.mathworks.com

http://www.simbrain.net/

PROJECT DATA is provided on the course web page.

This is a real life classification data with TYPE DE ROCHE (Rock Type) as a CLASS attribute. There are 98 records with 48 attributes and 6 classes.

Classes are:

C1 : R. Carbonatees AND R. Carbonatees impures

C2 : Pyrate

C3 : Charcopyrite

C4 : Galene

C5 : Spahlerite

C6 : Sediments terrigenes

Most important attributes (as determined by the expert) are: **S, Zn, Pb, Cu, CaO+MgO, CaO, MgO, Fe2O3**

This is a real life experimental data and it contains a lot of missing data (no value).

The project has to follow the following steps of **DM Process** to build different classifiers defined to three experiments.

S1: Data Preparation that includes attributes selection, cleaning the data, filling the missing values, etc... to build Project DATA - **PD**.

S2: Data preprocessing

1. For the Decision Trees **Descriptive Classifier** you use 2 methods of data discretization to the Project Data **PD** creating two data sets: **PD1** and **PD2**. Describe which methods you used.

2. For the Neural Network **Non -descriptive Classifier** use the Project DATA - **PD** and your tool method of normalization of your choice. Specify which.

Building Classifiers

For each sets of data **PD1**, **PD2** (for Decision Trees), and **PD** (for Neural Networks) perform the following **Experiments 1- 3**.

For each Experiment **compare** the resulting **Descriptive Classifiers** with each other and compare each **Descriptive Classifier** with the resulting **Non-Decsriptive Classifier**.

Experiments 1- 3

Experiment 1 : use the preprocessed data to perform a **full classification** (learning), i.e. build a classifier for all classes **C1- C6** simultaneously.

Experiment 2 : use all records to perform a **contrast classification** (contrast learning), i.e. a classification contrasting class **C1** with a class **notC1** that contains other classes.

Experiment 3 : repeat Experiments 1, 2 for all records with the **most important attributes** as defined by the expert.

Write a detailed Project Description with methods, motivations, results and submit via e-mail to TA and Professor.

It is a team project - the teams are the same is the for the Research Presentation.

RESEARCH PRESENTATION DESCRIPTION

Each presentation must consists of the following two parts.

Part 1 (40pts)

It is a **lecture type** 25 minutes long presentation (see description below).

Part 2 (20pts)

It is a short, 5 minutes presentation of a research paper, or an application (see description below)

Presentation Part 1 main goal is to **teach others** the material. It is a detailed, lecture type presentation. It can be based on, or extending the content of the book, book slides (if you need them come and copy from me), my slides, or any other sources.

Presenters have to put time and effort into **understanding the material**, present it slowly and be prepared to answer questions.

Remember that "I don't understand" is also an answer, but don't over-use it! The better answer is: "the book is not very clear, I think that it is ..., or I understood it as ...".

Presentation Part 2 is a presentation of a research paper, or a newest commercial application **connected with** the subject covered in the Presentation Part 1.

The structure of the **Presentation Part 2** is as follows:

1. If you present a paper you must include on your first slides authors names, title and place (journal, conference) where it was published and the date of the publication, or any other source of the paper you use.

You must PRINT a copy of the paper and put it in your **Presentation Folder**.

2. If you present a commercial application you must find relevant data about the application and include it in your **Presentation Folder**.

3. Each group member must present some part of the whole group work. The format of how you decide to do it is left to you as a group.

Presentation General Format

First slide must contain: Presentation Title, Presentation TEAM NUMBER, team members names and student IDs, Professor name, course number and course name.

Second slide must contain ALL sources you used for the your presentation. The book is included. In the case of the book the reference you have to put are title of the chapter, sections and pages numbers.

Third slide is an OVERVIEW of your presentation.

Remember to include a **source** of any picture, of any slides copied from a source, or any DIRECT citation on the bottom of each of your slides where it appears.

Presentation has to be given in **teams** of 4 students.

Presenters will be graded for the presentation skills, the content, organization, clarity, and amount of work put into research and preparation form and delivery.

Each member of the team has to present his/hers own well defined part and will be graded individually on this part as well as a part of overall evaluation of the group.

Presentations will be available on the course webpage for other students to help them to write their **final presentations reports**.

Of course students **should** attend the presentations to **learn** the material and evaluate the presentation delivery.

I will **collect** their **preliminary reports** (Part Four of the Final Report) written in class during the presentations.

Presentation Folder

Each team must give to Professor their **Presentation Folder** before they start the presentation.

The Presentation Folder must be labeled with students names, ID and Presentation TEAM number. It must contain the following.

1. A hard copy of the presentation (black and white in slide spread format)

2. PRINTED a copy of the paper, if yiou present a paper.

3. If you present a commercial application you must find relevant data about the application.

You receive 0-10pts for the organization and content of the Presentation Folder.

PRESENTATION SUBJECTS - students can find their own subjects; here are some possible subjects.

Data Warehouse and OLAP technology for Data Mining. (Chapter 3 of the Book)

Data Cube Computation and Data Generalization (Chapter 4 of the Book)

Statistical Methods 1: Statistical Prediction, Prediction by Regression, other purely statistical methods

Statistical Methods 2: Classification by Neural Networks

Statistical Methods 3: Bayesian Classification.

Statistical Methods 4: Cluster Analysis. A Categorization of major Clustering methods.

Evolutionary Computing: Genetic algorithms as optimization, Genetic algorithms as classification. Other evolutionary computing methods.

NEW ADVANCES in Data Mining, for example.

Deep Learning

Web Mining: an overview of methods and problems

Text Mining: an overview of methods and problems

Visualization and DM techniques

Natural Language Processing and DM techniques

FIND YOUR OWN subject and discuss it with the Professor.

FINAL REPORT

Each student has to write a report about 10 research presentations.

Report is DUE last DAY of classes, or any day before. Mail finished PARTS ONE - THREE to Professor and TA.

Print and bring PART FOUR to class. Fill it and give to Professor at the end of presentations.

Below of the format for presentations report. Make notes during students presentations keeping in mind what you have to include in them.

PRESENTATION REPORT FORMAT

PART ONE: Write your own general opinion about the talk, speakers, subject. Compare with at least 3 other talks.

PART TWO: Write your own short description of the **content** of the LECTURE part of the presentation.

PART THREE: Write your own short description of the **content** of the PAPER or APPLICATION PART part of the presentation.

Required Syllabi Statements: The University Senate has authorized that the following required statements appear in all teaching syllabi on the Stony Brook Campus.

Americans with Disabilities Act: If you have a physical, psychological, medical or learning disability that may impact your course work, please contact Disability Support Services, ECC(Educational Communications Center) Building, Room 128, (631)632-6748. They will determine with you what accommodations, if any, are necessary and appropriate. All information and documentation is confidential.

Academic Integrity: Each student must pursue his or her academic goals honestly and be personally accountable for all submitted work. Representing another person's work as your own is always wrong. Faculty is required to report any suspected instances of academic dishonesty to the Academic Judiciary. Faculty in the Health Sciences Center (School of Health Technology & Management, Nursing, Social Welfare, Dental Medicine) and School of Medicine are required to follow their school-specific procedures.