CS 533 -- Data Mining & Big Data Analysis
Fall 2015

Class Time: 1-1:50 p.m. MWF
Class Venue: ENGA 222
Course Website: http://www.cs.siu.edu/~dche/courses/CS533/

INSTRUCTOR

Dr. Dunren “Daren” Che, Professor
Office: Faner 2128
Office Hours: 11: 00 am - 1:00 pm MWF
Phone: 453-6046
Email: dche@cs.siu.edu (preferred!!)

PREREQUISITES

430 with a grade of C or better, plus enthusiasm and hard work.

COURSE DESCRIPTION

Prologue: Considering this is a graduate level course, the instructor may purposefully (as he usually does) add cutting-edge and/or emerging topics in Computer Science as supplements (These added topics are not required subjects for exams but are acceptable topics for students to choose to fulfill the research requirement of this course). At the present time, Cloud Computing, and Big Data are buzzwords, indicating exciting research frontiers in Computer Science, and will be added to this course.

The main theme of this course is on the concepts and techniques of data mining. Data mining, which is more generally referred to as knowledge discovery (from data repositories), has emerged as one of the most exciting fields in Computer Science. Data mining aims at finding useful (unknown) regularities in large data sets. Interest in the field is motivated by the continued growth of computerized data collections which are routinely kept by corporate organizations and commercial enterprises, and by the high potential value of the hidden patterns discovered in these data collections. For instance, bar code readers at supermarkets produce extensive amounts of data about purchases; and analysis of this data may reveal previously unknown, yet highly useful information about the shopping behavior of the customers, which can be utilized to promote profits of the corporate (data owner).

Data mining refers to a set of techniques that have been designed to efficiently find interesting pieces of information or knowledge hidden in large amounts of data. Association rules, for instance, are a class of patterns that tell which products tend to be purchased together. There is currently a large commercial interest in the area, both for the development of data mining software and for the offering of consulting services on data mining.

In this course, we explore how this interdisciplinary field brings together techniques from databases, statistics, machine learning, and information retrieval. We will discuss the main data mining methods currently used, including data cleaning and preparation, classification, clustering, and association rule mining, and possibly anomaly detection. Designing algorithms for these tasks is difficult because the input data sets are assumed to be very large, and the tasks may be very complex. This is especially true in the new era of “Big Data”.
Main Topics to be covered:

1. Introduction to Data Mining------------------------ 5 lectures
2. Introduction to Big Data and Big Data Analytics ---------    5 lectures
3. Data Cleaning/Transformation/Preparation ----------------    5 lectures
4. Association Rule Mining ----------------------------- 5 lectures
5. Classification/Prediction Techniques -----------------    5 lectures
6. Clustering Techniques ------------------------------- 5 lectures
7. Data Mining Applications ---------------------------- 4 lectures
8. Student Research Reports --------------------------- 6 lectures

TEXTBOOK

Introduction to Data Mining, by
Pang-Ning Tan, Michigan State University
Michael Steinbach, University of Minnesota
Vipin Kumar, University of Minnesota
ISBN-10: 0321321367
Publisher: Addison-Wesley
Copyright: 2006
Format: Cloth; 769 pp
Published: 05/02/2005
URL: http://www.pearsonhighered.com/educator/product/Introduction-to-Data-Mining/9780321321367.page

REQUIREMENT AND GRADING POLICY

A final letter grade will be assigned to each student based on the following 2 factors: a written exam (covering the core data mining topics only) and one research report (including presentation). All CS majors shall do research on a relevant topic (including Big Data and Cloud Computing topics) of their own choice and submit a research report; non-CS majors may instead opt to do a self-study on a relevant topic and submit a study report. Note that the self-study topic must be one that is not or not adequately covered by the instructor. The main difference between a research report and a study report is that with the former, you can’t just compile others materials – you have to have something of your own: new ideas on or original insight into an important issue, and/or experimental support or disproof of a technical method/approach. All reports are required to be presented to the class during the last couple of weeks of the semester.

Both factors carry equal weight, i.e.:

- Written exam------------------ 50%
- Research----------------------- 50% (of which 15% on presentation)

What is new with this course is that we will put relatively more attention on width vs. depth of the issues that we will address (This idea was motivated by a conversation with a former student, Patrick Brown). Graduate students (after a minimum of four years of formal training in the discipline) are supposed to have the ability to pick up any CS topic and study/deepen their understanding by themselves. Therefore, via this course I will give my priority to broadening students’ knowledge in the area rather than deepening their understanding on a limited select of technique issues.

This course encourages and promotes student research, purposefully leading students to a possible thesis research
topic if they haven’t chosen one yet. Basically, all students pick a topic (must inform the instructor by the end of the 4th week), collect and read at least 5 related papers, and produce a research/study report of about 10 pages and present it to the class (The report must be submitted to the instructor by email before Dec. 1st). The most successful example of the course research is that it will result in a publication at a good (peer-reviewed) conference or journal (The course had witnessed several such cases in the past few years).

Research reports will be graded using the criteria that are commonly used for reviewing academic journal/conference papers, e.g., readability and clarity of presentation, adequacy of literature review, adequacy of analysis of issues, and originality of the presented approach/ideas/methods, etc.

The final letter grade of students will be decided based on the following "standard" scales:

- A ---- above 90
- B ---- 80 to 89
- C ---- 70 to 80
- D ---- 60 to 69
- F ---- Less than 60

A critical advice: Start EARLY (!) because time is never plenty enough when you need it to produce a good research result.

A good attendance record may earn you an extra key credit that may make your final grade different (will explain this policy in detail in classroom).

A ROUGH SCHEDULE of Lectures

- Week 1 to Week 12: Data mining topics
- Week 13-15: Exam, student presentations, and final course wrap-up

ACADEMIC CHEATING PREVENTION

Simply put, do all assignments yourself; can seek help from each other through discussion and active learning but never copy (even a portion) from others. If identified, both sides involved will be punished according to department and/or university policies.

Alert: an overlooked cheating case with research/study reports – if you simply copy and paste from one or two papers without adequate digestion and reprocessing, it’s CHEATING and you will earn zero credit for this part of request!!!

FINAL NOTE REGARDING FAIRNESS OF GRADING

You can discuss any problem, but shall not try to negotiate with the instructor for a better grade and shall not try to challenge the instructor’s grading policy that has been set with the whole class in mind. When it is not a “black and white” type of problem in a test and your answer is not completely right. It is impossible to precisely quantify to what percent you are right and thus to ask for a commensurate portion of the points of that problem. As long as potential misunderstanding is clarified, you shall leave it to the instructor’s discretion. Be aware that after the same “grading scale” have been applied to all the students in the class, your request for a more generous score (even though might sound reasonable) simply means that a similar change needs be made to all other students with similar cases – a disaster that every instructor would try to avoid. The instructor must assure basic fairness to the whole class. Students shall not make any individual request that conflicts with this basic “fairness rule”.
University Required Syllabus Attachment: please click here