CSE 601 -- Data Mining and Bioinformatics
Course Syllabus

INSTRUCTORS	OFFICE HOURS
Prof. Aidong Zhang	349 Davis Hall
Phone: 645-4769	Saturdays: 3:30pm-4:30pm
Email: azhang@buffalo.edu	

Texts

- **Data Mining: Concepts and Techniques, 3rd ed.** Jiawei Han and Micheline Kamber, ISBN-13: 978-1-55860-901-3, Morgan Kaufmann Publishers.
- **Introduction to Data Mining.** Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Addison Wesley.
- **Data Warehousing.** Paulraj Ponniah. John Wiley & Sons, Inc.

References

- **Bioinformatics: Managing Scientific Data**. Zoe Lacroix and Terence Critchlow. 2003. Morgan Kaufmann Publishers.
- **Advanced Analysis of Gene Expression Microarray Data.** Aidong Zhang. ISBN 981-256-645-7. World Scientific Publishing Co.

Course Description

Main focus:

This course focuses on the fundamental techniques in data mining, including data warehousing, frequent pattern mining, clustering, classification, anomaly detection and feature selection methods. Specifically, we will cover the following topics:

- Data warehousing – model design
- Frequent pattern mining – association rules mining
- Clustering – partition-based, hierarchical-based, density-based approaches, spectral clustering
- Feature selection – dimensionality reduction
- Classification – decision-tree, Bayesian, rule-based, SVM, ensemble methods
- Anomaly detection – statistics-based, density-based, clustering-based
- Evaluation and validation of data mining results
- Correlation analysis – metrics and analysis
- Graph and network mining
- Visualization of patterns – mapping between high-dimensional data and low-dimensional data
- Multi-source information integration

Applications in Bioinformatics:

To demonstrate how data mining techniques are applied to various domains, we focus on the software systems design of bioinformatics, discussing the applications of data warehousing and data mining in biological and biomedical related fields. The class will discuss various software systems and provide insight that will help students gain a comprehensive understanding of the bioinformatics field. Projects will be designed based on these applications.

Focused project topics:

Project I –Databases/Data Warehouse Design

- Modeling: Conceptual and logical modeling of biomedical data
- System design of biomedical data warehouses
- Online Analytical Processing (OLAP) tools for biomedical data

Project II –Cluster Analysis

- Implement several clustering algorithms that partition data points into groups based on their similarity
- Implement parallel clustering algorithm on MapReduce
- Evaluate the clustering results using internal or external index

Project III –Classification

- Build classifiers that learn from training data and apply to test data to predict their class labels
- Select informative features that lead to a good classifier
- Evaluate the performance of the classification results

Homework Assignments

There will be three or four written homework assignments. You may work as a team of at most three people for homework and projects. Each team will have to present your results of the projects in class. Each project requires a term paper.

There will be class quiz but no formal written examinations. This class will emphasize on research and creative work.

Grading (subject to change)

Class participation – 5%
Quizzes -- 20%
Projects (3) -- 45%
Homework (3-4) -- 30%