10B22CI622: Data Mining

Course Credit: 4
Semester: VI

Introduction

Data Mining studies algorithms and computational paradigms that allow computers to find patterns and regularities in databases, perform prediction and forecasting, and generally improve their performance through interaction with data. It is currently regarded as the key element of a more general process called Knowledge Discovery that deals with extracting useful knowledge from raw data. The knowledge discovery process includes data selection, cleaning, coding, using different statistical, pattern recognition and machine learning techniques, and reporting and visualization of the generated structures.

This course will offer a comprehensive coverage of well known Data Mining Topics including classification, clustering and association rules. A number of specific algorithms and techniques under each category will be discussed. Methods for feature selection, dimensionality reduction and performance evaluation will also be covered. Students will learn and work with appropriate software tools and packages in the laboratory. They will be exposed to relevant Data Mining research.

Course Objectives (Post-conditions)

Knowledge objectives:

1. You will broaden your knolwedge of Information Systems
2. You will aware of Data Mining concepts
3. You will understand Data preprocessing task in data mining
4. You will gain practical experiennce on Visualization Techniques in data mining
5. You will learn what is Frequent Pattern Mining.
6. You will learn Predictive Data Mining tasks
7. You will become aware of Clustering and its usefulness.
8. You will learn concepts of outlier minng.
9. You will learn support vector machines and Regression
9. Introduction to WEKA tool for data mining tasks

Application objectives:

1. Designing and implementation of Principle Component Analysis algorithm for dimensionality reduction on given real data set (in MATLAB or R) OR
Implementing Bayesian, Naïve and Decision tree based classifier on given data sets (MATLAB or R).

Project involves following key tasks:

1. Preparing and preprocessing the data set
2. Apply three classification algorithms listed above

3. Compare results achieved by various classification algorithms using performance metrics.

4. Analyze and visualize results.

Expected Student Background (Preconditions)

For this course you need basic computing proficiency including some programming experience in a typical programming language, such as C, Java, or Python, knowledge of basic concepts of databases, artificial intelligence, and statistics.

Topics Outline:

S NO	Topics	Hrs
1	Overview of Data Mining	3
2	Data Preprocessing	4
3	Visualization Techniques	3
4	Dimensionality Reduction	5
5	Mining Frequent Patterns, Associations, and Correlations	8
6	Classification	9
7	Cluster Analysis	9
8	Support Vector Machines and Regression	5
9	Outlier detection	3
	Total	49

References

1. Introduction to Data Mining Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Pearson Education (Addison Wesley), 0-321-32136-7, 2006
2. Mininig Massive data sets Anand Rajaram, Jure Leskovec and Jeff Ullman Cambridge University Press
3. Data Mining Concepts and Techniques J. Han and M. Kamber Morgan Kaufmann, 2006, ISBN 1-55860-901-6
4. An Introduction to Information Retrieval, 2008 Cambridge UP.

Evaluation Scheme:

S.No	Examination	Marks
1	T-1	15
2	T-2	25
3	T-3	35
4	*Internal Marks	25

*Internal Marks Breakdown:

Assignments 9 marks (3x3)

Quizzes 12 marks (3x4)

Regularity 4 Marks