IT-772 : Data mining and modeling (DMM)

Pre-requisites: There is no additional pre-requisite for the course, as students having passed Pre Ph.D. semseter **I** exam are considered suitable for the course.

Objective: This course is aimed at training students in general principles of data mining, specially applied to biological data analysis.

Syllabus:

Introduction to Data Mining: Data Mining Techniques, Knowledge Representation Methods, Applications, Examples of sequence, PSSM data in biological systems.

Data preprocessing: Data cleaning transformation and reduction. Discretization and generating concept hierarchies

Attribute-oriented analysis: Attribute generalization, relevance and comparison, Statistical tests for assessment.

Data mining algorithms: Association rules, item sets, Generating item sets and rules efficiently, Correlation analysis. Classification, Basic learning tasks, Inferring rudimentary rules: 1R algorithm, Decision trees, Covering rules. Prediction, Statistical (Bayesian) classification, Bayesian networks, Instance-based methods (nearest neighbor), Linear models. Evaluating models. Estimating classifier accuracy (holdout, cross-validation, leave-one-out). Combining multiple models (bagging, boosting, stacking), Minimum Description Length Principle (MLD). overfitting, regularization

Clustering: Basic issues in clustering, Partitioning methods: k-means, expectation maximization (EM), Hierarchical methods: distance-based agglomerative and divisible clustering

Advanced techniques: Text mining: extracting attributes (keywords), structural approaches (parsing, soft parsing). Bayesian approach to classifying text. Web mining: classifying web pages, extracting knowledge from the web

Suggested readings:

1. Ian H. Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques (Second Edition), Morgan Kaufmann, 2005, 1SBN:0-12-088407-0

2. Introduction to Information Retrieval, Christopher D. Manning, Prabhakar Raghavan and Hinrich Schtitze, Cambridge University Press. 2008.

3. Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach and Vipin Kumar, Addison-Wesley, 2006