Kadi Sarva Vishwavidyalaya, Gandhinagar
MASTERS OF COMPUTER APPLICATION (MCA)
Year – III (Semester – V) (W.E.F. June 2015)
Subject Name: Data Warehousing & Data Mining (DWDM) – MCA-501

Sub Total Credit	Teaching scheme		Examination scheme				
	(per week)		MID	CEC	External		Total Marks
	Th	Pr	Th	Th	Th.	Pr.	
5	3	4	25	25	50	50	150

Course Description:

Data warehousing and data mining are two major areas of exploration for knowledge discovery in databases. These topics have gained great relevance especially in the 1990's and early 2000's with web data growing at an exponential rate. As more data is collected by businesses and scientific institutions alike, knowledge exploration techniques are needed to gain useful business intelligence. This course will cover a wide spectrum of industry standard techniques using widely available database and tools packages for knowledge discovery.

Data mining is for relatively unstructured data for which more sophisticated techniques are needed. The course aims to cover powerful data mining techniques including clustering, association rules, and classification. It then teaches high volume data processing mechanisms by building warehouse schemas such as snowflake, and star. OLAP query retrieval techniques are also introduced.

Learning Objectives

- To understand the need of Data Warehouses over Databases, and the difference between usage of operational and historical data repositories.
- To be able to differentiate between RDBMS schemas & Data Warehouse Schemas.
- To understand the concept of Analytical Processing (OLAP) and its similarities & differences with respect to Transaction Processing (OLTP).
- To conceptualize the architecture of a Data Warehouse and the need for pre-processing.
- To understand the need for Data Mining and advantages to the business world. The validating criteria for an outcome to be categorized as Data Mining result will be understood.
- To get a clear idea of various classes of Data Mining techniques, their need, scenarios (situations) and scope of their applicability.
- To learn the algorithms used for various types of Data Mining problems.

Pre-requisites: Knowledge of RDBMS and OLTP

Unit: 1 – Introduction to Data Warehousing, A Multi-dimensional Data Model & Schemas, OLAP Operations & Servers

- An overview and definition along with clear understanding of the four key-words appearing in the definition.
- Differences between Operational Database Systems and Data Warehouses; Difference between OLTP & OLAP
- Overview of Multi-dimensional Data Model, and the basic differentiation between "Fact" and "Dimension"; Multi-dimensional Cube
- Concept Hierarchies of "Dimensions" Parameters: Examples and the advantages
- Star, Snowflakes, and Fact Constellations Schemas for Multi-dimensional Databases
- Measures: Their Categorization and Computation

- Pre-computation of Cubes, Constraint on Storage Space, Possible Solutions
- OLAP Operations in Multi-dimensional Data Model: Roll-up, Drill-down, Slice & Dice, Pivot (Rotate)
- Indexing OLAP Data; Efficient Processing of OLAP Queries
- Type of OLAP Servers: ROLAP versus MOLAP versus HOLAP
- Metadata Repository

Data Warehouse Architecture; Further Development of Data Cube & OLAP Technology

- The Design of A Data Warehouse: A Business Analysis Framework; The Process of Data Warehouse Design
- A 3-Tier Data Warehouse Architecture; Enterprise Warehouse, Data mart, Virtual Warehouse
- Discovery-Driven Exploration of Data Cubes; Complex Aggregation at Multiple Granularity: Multi-feature Cubes
- Constrained Gradient Analysis of Data Cubes

Unit: 2 – Pre-processing

- The need for Pre-processing, Descriptive Data Summarization
- Data Cleaning: Missing Values, Noisy Data, Data Cleaning as a Process
- Data Integration & Transformation
- Data Cube Aggregation; Attribute Subset Selection
- Dimesionality Reduction: Basic Concepts only
- Numerosity Reduction: Regression & Log-linear Models, Histograms, Clustering, Sampling
- Data Dicretization & Concept Hierarchy Generation
- For Numerical Data: Binning, Histogram Analysis, Entropy-based Discretization, Interval Merging by x2 Analysis, Cluster Analysis, Discretization by Intuitive Partitioning
- For Categorical Data

Data Mining: Introduction

- An Overview; What is Data Mining; Data Mining – on What Kind of Data
- Data Mining Functionalities – What Kind of Patterns Can be Mined; Concept/Class Description: Characterization & Discrimination; Mining Frequent Patterns, Associations, and Correlations; Classification & Prediction; Cluster Analysis; Outlier Analysis
- Are All of the Patterns Interesting
- Classification of Data Mining Systems
- Data Mining Task Primitives
- Integration of a Data Mining System with a Database or Data Warehouse System
- Major Issues in Data Mining

Unit: 3 – Attribute-Oriented Induction: An Alternate Method for Data Generalization & Concept Description

- Attribute-Oriented Induction for Data Characterization, and Its Efficient Implementation; Presentation of the Derived Generalization
- Mining Class Comparisons: Discrimination between Different Classes
- Class Descriptions: Presentation of both Characterization & Comparison

Unit: 4 – Mining Frequent Patterns, Associations, and Correlations

- Basic Concepts: Market Basket Analysis; Frequent Itemsets, Closed Itemsets, and Association Rules; Frequent Pattern Mining: A Roadmap
- Apriori Algorithm: Finding Frequent Itemsets Using Candidate Generation; Generating Association Rules from Frequent Itemsets; Improving the Efficiency of Apriori
- From Association Mining to Correlation Analysis; Strong Rules Are Not Necessarily Interesting: An Example; From Association Analysis to Correlation Analysis

Unit: 5 – Classification & Prediction

- Introduction to Classification and Prediction; Basics of Supervised & Unsupervised Learning; Preparing the Data for Classification and Prediction; Comparing Classification and Prediction Methods
- Classification by Decision Tree Induction, Attribute Selection Measures; Tree Pruning; Scalability and Decision Tree Induction
- Rule-based Classification: Using IF-THEN Rules for Classification; Rule Extraction from a Decision Trees; Rule Induction Using a Sequential Covering Algorithm
- Bayesian Classification: Bayes' Theorem, Naïve Bayesian Classification; Bayesian Belief Networks
- An Overview of Other Classification Methods (2 Lectures)
- Prediction: Linear Regression; Non-linear Regression; Other Regression Models
- Classifier Accuracy and Error Measures: Classifier Accuracy Measures; Predictor Error Measures
- Evaluating the Accuracy of a Classifier or Predictor: Holdout Method and Random Subsampling; Cross Validation; Bootstrap
- Ensemble Methods – Increasing the Accuracy: Bagging; Boosting

Cluster Analysis

- Introduction to Cluster Analysis; Types of Data in Cluster Analysis; A Categorization of major Clustering Methods
- Partitioning Methods; Centroid-Based Technique: K-Means Method; Overview of Other Clustering Methods
- **An Overview of Other Clustering Methods (2 Lectures)**
- Outlier Analysis; Statistical Distribution-based Outlier Detection; Distance-based Outlier
- Detection; Density-based Outlier Detection; Deviation-based Outlier Detection

Chapter wise Coverage from the Text Books

Unit-1: 3.1, 3.1.1, 3.2, 3.2.1 to 3.2.6, 3.4.1 to 3.4.3, 3.3.4, 3.3.5, 3.3, 3.3.1, 3.3.2, 4.2.1 to 4.2.3

Unit-2: 2.1, 2.2, 2.2.1 to 2.2.3, 2.3.1 to 2.3.3, 2.4.1, 2.4.2, 2.5.1, 2.5.2, (Introductory Portion of 2.5.3), 2.5.4, 2.6, 2.6.1, 2.6.2, 1.1 to 1.3: 1.3.1 to 1.3.4, 1.4, 1.4.1 to 1.4.5, 1.5 to 1.9

Unit-3: 4.3.1 to 4.3.5

Unit-4: 5.1.1 to 5.1.3, 5.2.1 to 5.2.3, 5.4, 5.4.1, 5.4.2

Unit-5: 6.1, 6.2, 6.2.1, 6.2.2, 6.3, 6.3.1 to 6.3.4, 6.5, 6.5.1 to 6.5.3, 6.4, 6.4.1 to 6.4.3, 6.11, 6.11.1 to 6.11.3, 6.12,6.12.1, 6.12.2, 6.13, 6.13.1 to 6.13.3, 6.14, 6.14.1, 6.14.2, 7.1, 7.2, 7.2.1 to 7.2.5, 7.3, 7.4, 7.4.1, 7.11, 7.11.1 to 7.11.4

Accomplishment of the students after completing the course

- ✓ Ability to create a Star Schema for a given Data Warehousing requirements
- ✓ Ability to decide the number & levels of pre-computed Data Cubes, the corresponding Metadata, and the appropriate OLAP operation Warehouse
- ✓ Ability to apply pre-processing on existing operational & historical data for creation of Data
- ✓ Ability to apply Apriori algorithm for Association Mining
- ✓ Ability to apply Decision Tree and Bayesian algorithms for Classification
- ✓ Ability to mine Statistical Measures in large databases
- ✓ Ability to differentiate between Classification & Clustering, and similarly between Supervised Learning & Unsupervised Learning

Suggested Continuous Evaluation Components (CEC) Data Warehousing & Datamining

- ✓ Data Warehouse Applications: CRM; SCM; Banking sector; Insurance sector; Retail banking Industry case study, Hospital application.
- ✓ Design a data mart from scratch to store the credit history of customers of a bank. Use this credit profiling to process future loan applications.
- ✓ Design and build a Data Warehouse using bottom up approach titled 'Citizen Information System'. This should be able to serve the analytical needs of the various government departments and also provide a global integrated view.

Group Project

- ✓ Based on their collective work experience, each group should identify, and to the extent possible, execute a business intelligence project that relies on the data mining techniques we will cover in the class. The key tasks here are:
- ✓ To identify a business problem or a series of interesting questions that deal with either classification, prediction or clustering
- ✓ Identify sources of data that could potentially be useful in addressing your questions
- ✓ Pre-process – clean, validate, visualize your data
- ✓ Develop your model considering alternative techniques, selecting the most appropriate one in the process.
- ✓ Interpret your results, and write a final report including an executive summary of your findings. This will be due during the finals week.
- ✓ Prepare a 10-15 minute presentation for the last class meeting

Laboratory Exercise to be performed on WEKA using the given dataset

Association Rules:

1. Try to find association rules for car database. Does all the rules are good?

2.Modify the car so that all the classes have a uniform distribution.

Try to find association rules from modified car database. What happened to the rules now?

3.Try to find association rules for credit database.

Then remove the attribute "foreing_worker". What happened to the rules now?

4.Try to find association rules for one or more of the remaining databases. List the rules as per the lift ratio. Does all the rules are important?

Clustering:

1. Select Iris database and applied density-based clustering technique.

How is the distribution of data lookalike?

2.Performs the same operation with Centroid-Based clustering technique on following databases: wine or WDBC. What are the clusters lookalike?

3.Performs the same operation with the database sponge.

Are you able to interpret the results in this case?

Decision Tree and Basian Classification

1. Perform the following tests on whether database: generates a decision tree and a Bayesian Classifiers.

- Repeat network to solve the problem, and performs a validation with 10 folds.
    - Repeat the experiment validating the results on the training set itself.
    - Repeat the same test with iris database.
    - Justify the results.

2. Repeat once more the same tests with the contact database. Compare the results with the results of exercise 1, justify the classes.

  - ✓ Try different classifiers on whether database and compare the results of them.
  - ✓ Try different attributes to classify the database, use attribute selection method to select splitting attributes.