Syllabus

Savitribai Phule Pune University
Fourth Year of Computer Engineering (2015 Course)
Elective I

410244(D) : Data Mining and Warehousing

Teaching Scheme : TH : 03 Hours/Week	Credit 03	Examination Scheme : In-Sem (Paper) : 30 Marks End-Sem (Paper) : 70 Marks

Pre-requisites Courses

310242-Database Management Systems, 310244 - Information Systems and Engineering Economics

Companion Course : 410247- Laboratory Practice II

Course Objectives

- To understand the fundamentals of Data Mining.
- To identify the appropriateness and need of mining the data.
- To learn the preprocessing, mining and post processing of the data.
- To understand various methods, techniques and algorithms in data mining.

Course Outcomes

On completion of the course the student should be able to :

- Apply basic, intermediate and advanced techniques to mine the data.
- Analyze the output generated by the process of data mining.
- Explore the hidden patterns in the data.
- Optimize the mining process by choosing best data mining technique.

Course Contents

Unit I : Introduction **(08 Hours)**

Data Mining, Data Mining Task Primitives, Data : Data, Information and Knowledge; Attribute Types : Nominal, Binary, Ordinal and Numeric attributes, Discrete versus Continuous Attributes; Introduction to Data Preprocessing, Data Cleaning : Missing values, Noisy data; Data integration : Correlation analysis; transformation : Min-max normalization, z-score normalization and decimal scaling; data reduction : Data Cube Aggregation, Attribute Subset Selection, sampling; and Data Discretization : Binning, Histogram Analysis **(Refer chapter 1)**

Unit II : Data Warehouse **(08 Hours)**

Data Warehouse, Operational Database Systems and Data Warehouses (OLTP Vs OLAP), A Multidimensional Data Model: Data Cubes, Stars, Snowflakes, and Fact Constellations Schemas; OLAP Operations in the Multidimensional Data Model, Concept Hierarchies, Data Warehouse Architecture, The Process of Data Warehouse Design, A three-tier data warehousing architecture, Types of OLAP Servers : ROLAP versus MOLAP versus HOLAP. **(Refer chapter 2)**

Unit III : Measuring Data Similarity and Dissimilarity (08 Hours)

Measuring Data Similarity and Dissimilarity, Proximity Measures for Nominal Attributes and Binary Attributes, interval scaled; Dissimilarity of Numeric Data : Minskowski Distance, Euclidean distance and Manhattan distance; Proximity Measures for Categorical, Ordinal Attributes, Ratio scaled variables; Dissimilarity for Attributes of Mixed Types, Cosine Similarity. **(Refer chapter 3)**

Unit IV : Association Rules Mining (08 Hours)

Market basket Analysis, Frequent item set, Closed item set, Association Rules, a-priori Algorithm, Generating Association Rules from Frequent Item sets, Improving the Efficiency of a-priori, Mining Frequent Item sets without Candidate Generation : FP Growth Algorithm; Mining Various Kinds of Association Rules : Mining multilevel association rules, constraint based association rule mining, Meta rule-Guided Mining of Association Rules.

(Refer chapter 4)

Unit V : Classification (08 Hours)

Introduction to : Classification and Regression for Predictive Analysis, Decision Tree Induction, Rule-Based Classification : using IF-THEN Rules for Classification, Rule Induction Using a Sequential Covering Algorithm. Bayesian Belief Networks, Training Bayesian Belief Networks, Classification Using Frequent Patterns, Associative Classification, Lazy Learners-k-Nearest-Neighbor Classifiers, Case-Based Reasoning. **(Refer chapter 5)**

Unit VI : Multiclass Classification (08 Hours)

Multiclass Classification, Semi-Supervised Classification, Reinforcement learning, Systematic Learning, Wholistic learning and multi-perspective learning. Metrics for Evaluating Classifier Performance : Accuracy, Error Rate, precision, Recall, Sensitivity, Specificity; Evaluating the Accuracy of a Classifier : Holdout Method, Random Sub sampling and Cross-Validation. **(Refer chapter 6)**

❑❑❑

UNIT I

Chapter 1 : Introduction 1-1 to 1-21

Syllabus :

Data Mining, Data Mining Task Primitives, Data : Data, Information and Knowledge; Attribute Types : Nominal, Binary, Ordinal and Numeric attributes, Discrete versus Continuous Attributes; Introduction to Data Preprocessing, Data Cleaning : Missing values, Noisy data; Data integration : Correlation analysis; transformation: Min-max normalization, z-score normalization and decimal scaling; data reduction : Data Cube Aggregation, Attribute Subset Selection, sampling ; and Data Discretization : Binning, Histogram Analysis.

✓ **Syllabus Topic :** Data Mining 1-1

1.1 Data Mining 1-1

1.1.1 Applications of Data Mining 1-1

1.1.2 Challenges to Data Mining **(Oct. 16)** 1-2

1.1.3 KDD Process (Knowledge Discovery in Databases) **(Aug. 17)** 1-2

1.1.4 Architecture of a Typical Data Mining System 1-4

✓ **Syllabus Topic :** Data Mining Task Primitives 1-4

1.2 Data Mining Task Primitives 1-4

✓ **Syllabus Topic :** Data - Data, Information and Knowledge 1-6

1.3 Data : Data, Information and Knowledge 1-6

✓ **Syllabus Topic :** Attributes Types - Nominal, Binary, Ordinal and Numeric Attributes, Discrete Versus Continuous Attributes 1-6

1.4 Attributes Types 1-6

✓ **Syllabus Topic :** Introduction to Data Pre-processing 1-8

1.5 Introduction to Data Pre-processing 1-8

1.6 Different Forms of Data Pre-processing **(Oct. 16, Dec. 16)** 1-8

✓ **Syllabus Topic :** Data Cleaning 1-9

1.6.1 Data Cleaning 1-9

1.6.1(A) Steps in Data Cleansing 1-9

✓ **Syllabus Topic :** Missing Values 1-10

1.6.1(B) Missing Values **(May 16, Dec. 16, Dec. 17)** 1-10

✓ **Syllabus Topic :** Noisy Data 1-11

1.6.1(C) Noisy Data 1-11

1.6.1(D) Inconsistent Data 1-13

✓ **Syllabus Topic :** Data Integration 1-13

1.6.2 Introduction to Data Integration 1-13

1.6.2(A) Entity Identification Problem 1-14

✓ **Syllabus Topic :** Correlation Analysis 1-14

1.6.2(B) Redundancy and Correlation Analysis 1-14

✓ **Syllabus Topic :** Data Transformation and Data Discretization - Min-max Normalization, Z-Score Normalization and Decimal Scaling 1-15

1.6.3 Data Transformation and Data Discretization 1-15

1.6.3(A) Data Transformation 1-15

1.6.3(B) Data Discretization 1-15

1.6.3(C) Data Transformation by Normalization **(May 17)** 1-15

✓ **Syllabus Topic :** Discretization by Binning 1-16

1.6.3(D) Discretization by Binning 1-16

✓ **Syllabus Topic :** Discretization by Histogram Analysis 1-17

1.6.3(E) Discretization by Histogram Analysis 1-17

✓ **Syllabus Topic :** Data Reduction 1-17

1.6.4 Data Reduction 1-17

1.6.4(A) Need for Data Reduction 1-17

1.6.4(B) Data Reduction Technique 1-18

✓ **Syllabus Topic :** Data Cube Aggregation, Attribute Subset Selection 1-18

4 1.6.4(B).1 Data Cube Aggregation 1-18

1.6.4(B).2 Dimensionality Reduction **(May 16, Dec. 17)** 1-18

1.6.4(B).3 Data Compression 1-19

✓ **Syllabus Topic :** Sampling 1-20

1.6.4(B).4 Numerosity Reduction 1-20

1.7 Solved University Questions and Answers 1-21

UNIT II

Chapter 2 : Data Warehouse 2-1 to 2-35

Syllabus :

Data Warehouse, Operational Database Systems and Data Warehouses (OLTP Vs OLAP), A Multidimensional Data Model : Data Cubes, Stars, Snowflakes, and Fact Constellations Schemas; OLAP Operations in the Multidimensional Data Model, Concept Hierarchies, Data Warehouse Architecture, The Process of Data Warehouse Design, A three-tier data warehousing architecture, Types of OLAP Servers : ROLAP versus MOLAP versus HOLAP.

✓ **Syllabus Topic :** Data Warehouse 2-1
2.1 Data Warehouse 2-1
2.1.1 Features of Data Warehouse 2-2
✓ **Syllabus Topic :** Operational Database Systems and Data Warehouses (OLTP Vs OLAP) 2-2
2.2 Operational Database Systems and Data Warehouses (OLTP Vs OLAP) 2-2
2.2.1 Why are Operational Systems not Suitable for Providing Strategic Information? 2-2
2.2.2 OLAP Vs OLTP 2-3
✓ **Syllabus Topic :** A Multidimensional Data Model 2-4
2.3 A Multidimensional Data Model 2-4
2.3.1 What is Dimensional Modeling ? 2-4
✓ **Syllabus Topic :** Data Cubes 2-4
2.3.2 Data Cubes 2-4
✓ **Syllabus Topic :** Stars Schemas 2-4
2.3.3 Star Schema 2-4
✓ **Syllabus Topic :** The Snowflake Schema 2-5
2.3.4 The Snowflake Schema 2-5
2.3.5 Star Flake Schema 2-6
2.3.6 Differentiate between Star Schema and Snowflake Schema 2-6
2.3.7 Factless Fact Table 2-6
✓ **Syllabus Topic :** Fact Constellation Schema or Families of Star 2-7
2.3.8 Fact Constellation Schema or Families of Star 2-7
2.3.9 Examples on Star Schema and Snowflake Schema 2-7
✓ **Syllabus Topic :** OLAP Operations in the Multidimensional Data Model 2-21
2.4 OLAP Operations in the Multidimensional Data Model 2-21
✓ **Syllabus Topic :** Concept Hierarchies 2-24
2.5 Concept Hierarchies 2-24
✓ **Syllabus Topic :** Data Warehouse Architecture 2-24
2.6 Data Warehouse Architecture 2-24
✓ **Syllabus Topic :** The Process of Data Warehouse Design 2-27
2.7 The Process of Data Warehouse Design 2-27
2.8 Data Warehousing Design Strategies or Approaches for Building a Data Warehouse 2-27
2.8.1 The Top Down Approach : The Dependent Data Mart Structure 2-27
2.8.2 The Bottom-Up Approach : The Data Warehouse Bus Structure 2-28
2.8.3 Hybrid Approach 2-29
2.8.4 Federated Approach 2-29
2.8.5 A Practical Approach 2-29
✓ **Syllabus Topic :** A Three-Tier Data Warehousing Architecture 2-30
2.9 A Three-Tier Data Warehousing Architecture 2-30
2.9.1 Data Warehouse and Data Marts 2-30
✓ **Syllabus Topic :** Types of OLAP Servers : ROLAP versus MOLAP versus HOLAP 2-31
2.10 Types of OLAP Servers : ROLAP versus MOLAP versus HOLAP 2-31
2.10.1 MOLAP 2-31

2.10.2 ROLAP 2-32

2.10.3 HOLAP 2-33

2.10.4 DOLAP 2-33

2.11 Examples of OLAP 2-33

UNIT III

Chapter 3 : Measuring Data Similarity and Dissimilarity 3-1 to 3-7

Syllabus :

Measuring Data Similarity and Dissimilarity, Proximity Measures for Nominal Attributes and Binary Attributes, interval scaled; Dissimilarity of Numeric Data : Minkowski Distance, Euclidean distance and Manhattan distance; Proximity Measures for Categorical, Ordinal Attributes, Ratio scaled variables; Dissimilarity for Attributes of Mixed Types, Cosine Similarity.

✓ **Syllabus Topic :** Measuring Data Similarity and Dissimilarity 3-1

3.1 Measuring Data Similarity and Dissimilarity 3-1

3.1.1 Data Matrix versus Dissimilarity Matrix 3-1

✓ **Syllabus Topic :** Proximity Measures for Nominal Attributes and Binary Attributes, Interval Scaled 3-2

3.2 Proximity Measures for Nominal Attributes and Binary Attributes, Interval Scaled 3-2

3.2.1 Proximity Measures for Nominal Attributes 3-2

3.2.2 Proximity Measures for Binary Attributes 3-2

3.2.3 Interval Scaled **(May 17)** 3-3

✓ **Syllabus Topic :** Dissimilarity of Numeric Data : Minkowski Distance, Euclidean Distance and Manhattan Distance 3-4

3.3 Dissimilarity of Numeric Data : Minkowski Distance, Euclidean Distance and Manhattan Distance 3-4

✓ **Syllabus Topic :** Proximity Measures for Categorical, Ordinal Attributes, Ratio Scaled Variables 3-5

3.4 Proximity Measures for Categorical, Ordinal Attributes, Ratio Scaled Variables 3-5

3.4.1 Categorical Attributes 3-5

3.4.2 Ordinal Attributes 3-5

3.4.3 Ratio Scaled Attributes 3-5

3.4.4 Discrete Versus Continuous Attributes 3-6

✓ **Syllabus Topic :** Dissimilarity for Attributes of Mixed Types 3-6

3.5 Dissimilarity for Attributes of Mixed Types 3-6

✓ **Syllabus Topic :** Cosine Similarity 3-6

3.6 Cosine Similarity 3-6

UNIT IV

Chapter 4 : Association Rules Mining 4-1 to 4-30

Syllabus :

Market basket Analysis, Frequent item set, Closed item set, Association Rules, a-priori Algorithm, Generating Association Rules from Frequent Item sets, Improving the Efficiency of a-priori, Mining Frequent Item sets without Candidate Generation : FP Growth Algorithm; Mining Various Kinds of Association Rules : Mining multilevel association rules, constraint based association rule mining, Meta rule-Guided Mining of Association Rules.

✓ **Syllabus Topic :** Market Basket Analysis 4-1

4.1 Market Basket Analysis 4-1

4.1.1 What is Market Basket Analysis? 4-1

4.1.2 How is it Used ? 4-1

Grocer shopker

4.1.3 Applications of Market Basket Analysis **(Dec. 17)** 4-2

✓ **Syllabus Topic :** Frequent Itemsets 4-2

4.2 Frequent Itemsets 4-2

✓ **Syllabus Topic :** Closed Itemsets 4-2

4.3 Closed Itemsets **(May 16, Dec. 16, Aug. 17)** 4-2

✓ **Syllabus Topic :** Association Rules 4-3

4.4 Association Rules strong 4-3

4.4.1 Finding the Large Itemsets 4-3

4.4.2 Frequent Pattern Mining 4-3

4.4.3 Efficient and Scalable Frequent Itemset Mining Method 4-4

✓ **Syllabus Topic :** A-priori Algorithm....................4-4

4.5 A-priori Algorithm....................4-4

4.5.1 Advantages and Disadvantages of Apriori Algorithm....................4-5

✓ **Syllabus Topic :** Generating Association Rules from Frequent Item Sets....................4-5

4.6 Generating Association Rules from Frequent Item Sets....................4-5

✓ **Syllabus Topic :** Improving the Efficiency of a-priori....................4-5

4.7 Improving the Efficiency of a-priori....................4-5

4.8 Solved Example on Apriori Algorithm....................4-6

✓ **Syllabus Topic :** Mining Frequent Item sets without Candidate Generation : FP Growth Algorithm....................4-20

4.9 Mining Frequent Item sets without Candidate Generation : FP Growth Algorithm....................4-20

4.9.1 FP-Tree Algorithm....................4-20

4.9.2 FP-Tree Size....................4-21

4.9.3 Example of FP Tree....................4-21

4.9.4 Mining Frequent Patterns from FP Tree....................4-23

4.9.5 Benefits of the FP-Tree Structure....................4-27

✓ **Syllabus Topic :** Mining Various Kinds of Association Rules....................4-27

4.10 Mining Various Kinds of Association Rules....................4-27

4.10.1 Mining Multilevel Association Rules....................4-27

✓ **Syllabus Topic :** Constraint based Association Rule Mining....................4-29

4.10.2 Constraint based Association Rule Mining....................4-29

✓ **Syllabus Topic :** Metarule-Guided Mining of Association Rule....................4-29

4.10.3 Metarule-Guided Mining of Association Rule....................4-29

4.11 Solved University Question and Answer....................4-29

UNIT V

Chapter 5 : Classification 5-1 to 5-34

Syllabus :

Introduction to : Classification and Regression for Predictive Analysis, Decision Tree Induction, Rule-Based Classification : using IF-THEN Rules for Classification, Rule Induction Using a Sequential Covering Algorithm. Bayesian Belief Networks, Training Bayesian Belief Networks, Classification Using Frequent Patterns, Associative Classification, Lazy Learners-k-Nearest-Neighbor Classifiers, Case-Based Reasoning.

✓ **Syllabus Topic :** Introduction to : Classification and Regression for Predictive Analysis....................5-1

5.1 Introduction to : Classification and Regression for Predictive Analysis....................5-1

5.1.1 Classification is a Two Step Process....................5-1

5.1.2 Difference between Classification and Prediction ...5-3

5.1.3 Issues Regarding Classification and Prediction....................5-3

5.1.4 Regression....................5-3

✓ **Syllabus Topic :** Decision Tree Induction....................5-5

5.2 Decision Tree Induction Classification Methods....................5-5

5.2.1 Appropriate Problems for Decision Tree Learning......5-5

5.2.2 Decision Tree Representation....................5-5

5.2.3 Algorithm for Inducing a Decision Tree....................5-6

5.2.4 Tree Pruning....................5-7

5.2.5 Examples of ID3....................5-7

✓ **Syllabus Topic :** Rule-Based Classification : using IF-THEN Rules for Classification....................5-27

5.3 Rule-Based Classification : using IF-THEN Rules for Classification....................5-27

5.3.1 Rule Coverage and Accuracy....................5-28

5.3.2 Characteristics of Rule-Based Classifier....................5-28

✓ **Syllabus Topic :** Rule Induction Using a Sequential Covering Algorithm....................5-29

5.4 Rule Induction Using a Sequential Covering Algorithm....................5-29

✓ **Syllabus Topic :** Bayesian Belief Networks 5-29

5.5 Bayesian Belief Networks 5-29

✓ **Syllabus Topic :** Training Bayesian Belief Networks 31

5.6 Training Bayesian Belief Networks 31

✓ **Syllabus Topic :** Classification Using Frequent Patterns : Associative Classification 5-31

5.7 Classification Using Frequent Patterns : Associative Classification 5-31

5.7.1 CBA 5-32

5.7.2 CMAR 5-32

✓ **Syllabus Topic :** Lazy Learners 5-32

5.8 Lazy Learners : (or Learning from your Neighbors) 5-32

✓ **Syllabus Topic :** K-Nearest-Neighbor Classifiers 5-32

5.8.1 K-Nearest-Neighbor Classifiers 5-32

5.8.2 CBR (Case Based Reasoning) 5-33

UNIT VI

Chapter 6 : Multiclass Classification 6-1 to 6-8

Syllabus :

Multiclass Classification, Semi-Supervised Classification, Reinforcement learning, Systematic, Learning, Wholistic learning and multi-perspective learning. Metrics for Evaluating Classifier Performance : Accuracy, Error Rate, precision, Recall, Sensitivity, Specificity; Evaluating the Accuracy of a Classifier : Holdout Method, Random Sub sampling and Cross-Validation.

✓ **Syllabus Topic :** Multiclass Classification 6-1

6.1 Multiclass Classification 6-1

6.1.1 Introduction to Multiclass Classification 6-1

✓ **Syllabus Topic :** Semi-Supervised Classification 6-1

6.2 Semi-Supervised Classification 6-1

✓ **Syllabus Topic :** Reinforcement Learning 6-2

6.3 Reinforcement Learning **(Dec. 16, May 17)** 6-2

6.3.1 Introduction to Reinforcement 6-2

6.3.2 Elements of Reinforcement Learning 6-2

6.3.3 Reinforcement Function and Environment Function 6-3

✓ **Syllabus Topic :** Wholistic Learning 6-3

6.3.4 Whole System Learning 6-3

✓ **Syllabus Topic :** Systematic Learning 6-3

6.4 Systematic Learning 6-3

✓ **Syllabus Topic :** Multi-Perspective Learning 6-4

6.5 Multi-Perspective Decision Making for Big Data and Multi-Perspective Learning for Big Data **(Dec. 15, May 16, Dec. 16, May 17, Dec. 17)** 6-4

6.5.1 Fundamental of Multi-perspective Decision Making and Multi-perspective Learning 6-4

6.5.2 Influence Diagram 6-4

✓ **Syllabus Topic :** Metrics for Evaluating Classifier Performance : Accuracy, Error Rate, Precision, Recall, Sensitivity, Specificity 6-5

6.6 Model Evaluation and Selection 6-5

6.6.1 Accuracy and Error Measures 6-5

✓ **Syllabus Topic :** Evaluating the Accuracy of a Classifier : Holdout Method 6-6

6.6.2 Holdout 6-6

✓ **Syllabus Topic :** Random Sub-sampling 6-7

6.6.3 Random Sub-sampling 6-7

✓ **Syllabus Topic :** Cross-Validation (CV) 6-7

6.6.4 Cross-Validation (CV) 6-7

6.7 University Questions and Answers 6-7

□□□