10-11	Digital signatures, Steganography and Digital Watermarking
12-13	Malicious Software: Types of malwares (viruses, worms, trojan horse, rootkits, bots), Memory exploits - Buffer overflow, Integer overflow
14-15	Security in Internet-of-Things, Security implications, Mobile device security - threats and strategies, Cyberlaws

Assessment Methods

Written tests, assignments, quizzes, presentations as announced by the instructor in the class.

Keywords

Security mechanisms, private and public key cryptography, malware detection, security in IoT.

Data Mining (BHCS17B) Discipline Specific Elective - (DSE)

Credit: 06

Course Objective

This course introduces data mining techniques and enables students to apply these techniques on real-life datasets. The course focuses on three main data mining techniques: Classification, Clustering and Association Rule Mining tasks.

Course Learning Outcomes

On successful completion of the course, students will be able to do following:

1. Pre-process the data, and perform cleaning and transformation.
2. Apply suitable classification algorithm to train the classifier and evaluate its performance.
3. Apply appropriate clustering algorithm to cluster data and evaluate clustering quality
4. Use association rule mining algorithms and generate frequent item-sets and association rules

Detailed Syllabus

Unit 1

Introduction to Data Mining - Applications of data mining, data mining tasks, motivation and challenges, types of data attributes and measurements, data quality.

Data Pre-processing - aggregation, sampling, dimensionality reduction, Feature Subset Selection, Feature Creation, Discretization and Binarization, Variable Transformation.

Unit 2

Classification: Basic Concepts, Decision Tree Classifier: Decision tree algorithm, attribute selection measures, Nearest Neighbour Classifier, Bayes Theorem and Naive Bayes Classifier, **Model Evaluation:** Holdout Method, Random Sub Sampling, Cross-Validation, evaluation metrics, confusion matrix.

Unit 3

Association rule mining: Transaction data-set, Frequent Itemset, Support measure, Apriori Principle, Apriori Algorithm, Computational Complexity, Rule Generation, Confidence of association rule.

Unit 4

Cluster Analysis: Basic Concepts, Different Types of Clustering Methods, Different Types of Clusters, K-means: The Basic K-means Algorithm, Strengths and Weaknesses of K-means algorithm, Agglomerative Hierarchical Clustering: Basic Algorithm, Proximity between clusters, DBSCAN: The DBSCAN Algorithm, Strengths and Weaknesses.

Practical

Section 1: Preprocessing

Q1. Create a file "people.txt" with the following data:

Age	agegroup	height	status	yearsmarried
21	adult	6.0	single	-1
2	child	3	married	0
18	adult	5.7	married	20
221	elderly	5	widowed	2
34	child	-7	married	3

i) Read the data from the file "*people.txt".*

ii) Create a ruleset *E* that contain rules to check for the following conditions:

1. The age should be in the range 0-150.
2. The age should be greater than yearsmarried.

3. The status should be married or single or widowed.
4. If age is less than 18 the agegroup should be child, if age is between 18 and 65 the agegroup should be adult, if age is more than 65 the agegroup should be elderly.

iii) Check whether ruleset *E* is violated by the data in the file people.txt.

iv) Summarize the results obtained in part (iii)

v) Visualize the results obtained in part (iii)

Q2. Perform the following preprocessing tasks on the `dirty_iris` dataset[ii].

1. Calculate the number and percentage of observations that are complete.
2. Replace all the special values in data with NA.
3. Define these rules in a separate text file and read them.
   (Use editfile function in R (package editrules). Use similar function in Python).
   Print the resulting constraint object.
   – Species should be one of the following values: setosa, versicolor or virginica.
   – All measured numerical properties of an iris should be positive.
   – The petal length of an iris is at least 2 times its petal width.
   – The sepal length of an iris cannot exceed 30 cm.
   – The sepals of an iris are longer than its petals.
4. Determine how often each rule is broken (violatedEdits). Also summarize and plot the result.
   Find outliers in sepal length using boxplot and boxplot.stats

Q3. Load the data from wine dataset. Check whether all attributes are standardized or not (mean is 0 and standard deviation is 1). If not, standardize the attributes. Do the same with Iris dataset.

Section 2: Data Mining Techniques

Run following algorithms on 2 real datasets and use appropriate evaluation measures to compute correctness of obtained patterns:

Q4. Run Apriori algorithm to find frequent itemsets and association rules

4.1 Use minimum support as 50% and minimum confidence as 75%

4.2 Use minimum support as 60% and minimum confidence as 60 %

Q5. Use Naive bayes, K-nearest, and Decision tree classification algorithms and build classifiers. Divide the data set into training and test set. Compare the accuracy of the different classifiers under the following situations:

5.1 a) Training set = 75% Test set = 25%

b) Training set = 66.6% (2/3rd of total), Test set = 33.3%

5.2 Training set is chosen by i) hold out method ii) Random subsampling iii) Cross-Validation. Compare the accuracy of the classifiers obtained.

5.3 Data is scaled to standard format.

Q6. Use Simple Kmeans, DBScan, Hierachical clustering algorithms for clustering. Compare the performance of clusters by changing the parameters involved in the algorithms.

Recommended Datasets for DataMining practicals

1. UCI Machine Learning repository.
2. KDD Datasets
3. Open data platform, Government of India (https://data.gov.in/)

References

1. Han, J., Kamber, M.,& Jian,P. (2011). *Data Mining: Concepts and Techniques*. 3rd edition. Morgan Kaufmann

2. Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining*. 1st Edition. Pearson Education.

Additional Resources

1. Gupta, G. K. (2006). *Introduction to Data Mining with Case Studies*. Prentice-Hall of India.

2. Hand, D., & Mannila, H. & Smyth, P. (2006). *Principles of Data Mining*. Prentice-Hall of India.

3. Pujari, A. (2008). *Data Mining Techniques*. 2nd edition. Universities Press.

Course Teaching Learning Process

- Use of ICT tools in conjunction with traditional class-room teaching methods
- Interactive sessions
- Class discussions

Tentative weekly teaching plan is as follows:

Week	Content
1	Introduction to Data Mining , Challenges , Data Mining Origins, Data Mining Tasks, Applications
2-3	Types of data, Data Quality, Data Pre-processing, Measures of similarity and dissimilarity
5-8	Classification - Preliminaries, General Approach to Solving a Classification Problem, Decision Tree Induction , Evaluating the Performance of a Classifier
8-9	Rule Based Classifier , Nearest Neighbor Classifiers, Bayesian Classifiers
10-11	Association Rules -Problem definition, Frequent item-set generation (Apriori algorithm), Rule generation
11-12	Clustering - Basic concepts of clustering analysis, K-Means
13-14	Agglomerative Hierarchical Clustering, DBSCAN
15	Quality of clustering

Assessment Methods

Written tests, assignments, quizzes, presentations as announced by the instructor in the class.

Keywords

data mining, classifiers, data pre-processing, metrics.

Advanced Algorithms (BHCS17C) Discipline Specific Elective - (DSE)

Credit: 06

Course Objective