Lab Manual

IT336 Data Mining Lab

SYLLABUS
List of Programs as Assignments:
Q1. Build a Data Warehouse and Explore WEKA tool.
Q2. Demonstration of preprocessing on various datasets. – discretize, remove noise, remove feature
Q3. Demonstration of Association rule process on dataset using apriori algorithm.
Q4. Demonstrate performance of classification on various data sets.
Q5. Demonstrate performance of clustering on various data sets.
Q6. Demonstrate performance of Regression on various data sets
Q7. Implement following algorithms for various datasets
A. Apriori Algorithm.
B. FP-Growth Algorithm.
C. K-means clustering.
Q8. Implement Bayesian Classification for various datasets
Q9 Implement Decision Tree for various datasets.
Q10. Implement Support Vector Machines.
Q11 Applications of classification for web mining.
Q12. Case Study on Text Mining or any commercial application
Books recommended:
Text Books :
1. Jiawei Han & Micheline Kamber - Data Mining Concepts & Techniques Publisher Harcout India. Private Limited.
Reference Books :
1. G.K. Gupta – Introduction to Data Mining with case Studies, PHI, New Delhi – 2006.
2. A. Berson& S.J. Smith – Data Warehousing Data Mining, COLAP, TMH, New Delhi – 2004.
3. H.M. Dunham & S. Sridhar – Data Mining, Pearson Education, New Delhi, 2006.

Table of Contents

Lab1-2 Planning Data Analysis and Mining Process (using Weka)

Objectives:

O1. Explore the Weka tool to identify and characterize a dataset; practice some data preprocessing techniques; and explore the visualization and analysis like identifying trends etc.

- Datasets: student.arff, labor.arff

a. You can select a dataset from a list of publicly available datasets at UCI Machine Learning Repository or the datasets section at Weka. You are also welcome to explore datasets on your own from other sources.

b. Briefly describe what the dataset is about and size of the dataset (e.g. number of tables, number of instances and attributes, etc.)

c. Discuss potential data mining applications for the dataset. Name one or two types of data mining (classification, clustering, etc.) you think would be relevant and discuss the potential mining results. E.g. if you think clustering is relevant, describe what a likely cluster might contain and what the real-world meaning would be.

O2. Demonstration of preprocessing on above datasets using Weka.

Your task for this assignment is to identify and characterize a dataset; practice some data preprocessing techniques; and explore the visualization and data preprocessing functionalities of Weka, an open source data mining toolkit developed in Java.

[Selecting or filtering attributes] [Removing an attribute] [Discretization]

a. Select one attribute and discuss appropriate measures of the central tendency and dispersion for the attribute. Use a subset of the attribute values (of your own choice) from the dataset and compute the mean, median, mode, range, quartiles, and variance for the attribute.

b. Discuss data quality issues of the dataset. Are there (potential) problems with certain data attributes? What would be appropriate responses to these quality issues?

c. Discuss one or two data preprocessing techniques that are likely required for the dataset. e.g. Is data smoothing or data reduction required and what would be an appropriate technique. Select one attribute and use a subset of attribute values to do the following: 1) partition them into appropriate number of bins by equal-frequency as well as equal-width partitioning, 2) use smoothing by bin means to smooth the data based on the above partitioning, 3) normalize the attribute based on min-max normalization and z-score normalization. Comment on which method you would prefer to use for partitioning, smoothing, and normalization for the given attribute.

[Guided Exercise]

[ref- http://www.apgcm.edu.in/images/data-mining-lab-manual.pdf, https://www.sircrrengg.ac.in/images/newsletter/ITMATERIALS/DMLab.pdf]

G1. Demonstration of preprocessing on dataset student.arff

i. Load data and list the attributed with type and basic statistics

ii. Are there (potential) problems (missing values, outliers) with certain data attributes? What would be appropriate responses to these quality issues?

iii. Selecting or filtering attributes – a) Add attribute (weka.filters.unsupervised.attribute.Add) b)Remove an attribute (“weka.filters.unsupervised.attribute.remove”): filter and remove R-7, and Save the new working relation as an arff file c) Convert string attribute to nominal (weka.filters.unsupervised.attribute.StringToNominal)

iv. Discretization – a) divide the values of age attribute into three bins (intervals). (weka.filters.unsupervised.attribute.Discretize). b) use smoothing by bin means to smooth the data based on the above partitioning

v. Normalize the attribute based on min-max normalization and z-score normalization. (weka.filters.unsupervised.attribute.Normalize) Comment on which method you would prefer to use for partitioning, smoothing, and normalization for the given attribute.

G2. Perform Decision Tree Classification using suitable attributes. Repeat decision tree classification with different set of attributes. Is there any effect on accuracy?

[Assignment]

A1. Perform preprocessing on dataset labor.arff [https://mlritm.ac.in/assets/cse/cse_lab_manuals/R20_cse_manuals/DATA%20MINING%20LAB%20Manual.pdf]

A2. Data Mining Refer dataset - **The German Credit Data**: dataset consisting of 1000 actual cases collected in Germany credit dataset (original) Excel spread sheet version of the German credit data (Down load from web)

1. List all the categorical (or nominal) attributes and the real-valued attributes separately.
2. What attributes do you think might be crucial in making the credit assessment? Come up with some simple rules in plain English using your selected attributes.

Hint- 1. For each attribute of Germen data set, a. Analyze the values of attribute. b. Find attribute, which can be used for making decision on credit. 2. Form sample rules on selected attribute to classify the customer as good. 3. Form the sample rules on selected attribute to classify the customer as bad.

3. One type of model that you can create is a Decision Tree -train a Decision Tree using the complete dataset as the training data. Report the model obtained after training. What % of examples can you classify correctly?

Hint - Create a decision tree by using J48 Technique for the complete dataset as the training data in Weka Explorer. 1. Open German data set arff file in Weka Explorer. 2. Select classifier tab, choose J48 decision tree and select training data set from test data option. 3. Start classification.

4. Is testing on the training set as you did above a good idea? Why or Why not?
5. One approach for solving the problem encountered in the previous question is using cross-validation? Describe what is cross -validation briefly. Train a Decision Tree again using cross -validation and report your results. Does your accuracy increase/decrease? Why?
6. Check to see if the data shows a bias against "foreign workers" (attribute 20), or "personal-status" (attribute 9). One way to do this (perhaps rather simple minded) is to remove these attributes from the dataset and see if the decision tree created in those cases is significantly different from the full dataset case which you have already done. To remove an attribute you can use the preprocess tab in Weka's GUI Explorer. Did removing these attributes have any significant effect? Discuss.
7. Another question might be, do you really need to input so many attributes to get good results? Maybe only a few would do. For example, you could try just having attributes 2, 3, 5, 7, 10, 17 (and 21, the class attribute (naturally)). Try out some combinations. (removed two attributes in problem 7.
8. Sometimes, the cost of rejecting an applicant who actually has a good credit (case 1) might be higher than accepting an applicant who has bad credit (case 2). Instead of counting the misclassifications equally in both cases, give a higher cost to the first case (say cost 5) and lower cost to the second case. You can do this by using a cost matrix in Weka. Train your Decision Tree again and report the Decision Tree and cross -validation results. Are they significantly different from results obtained in problem 6 (using equal cost)?
9. How can you convert a Decision Trees into "if –then -else rules". Make up your own small Decision Tree consisting of 2 - 3 levels and convert it into a set of rules.

Lab 3 Basic Programming constructs in R

[Guided Exercise]

G3. Basic programming in R using arrays, list, looping and conditional control, nd-array using numpy

[Assignment]

A3. Write R programs for following

1. For assigning a grade based on percentage of marks 90% or more "A Grade"; 80% or more "B Grade"; 70% or more "C Grade"; 60% or more "D Grade"; less than 60% "F Grade"
2. Create a vector of 10 equal spaced values in the range 1:20, and print the following : i) 2nd element, ii) 2nd last element, iii) **leaving out the 2nd element** iv) elements at indices 2to5 v) **elements at indices { 2, 5, 8 }**
3. Enter two matrices and print their sum i) using R function ii) using looping in R

Lab 4 Data handling in R

G4. Loading and manipulating datafiles using dataframes in Pandas

A4. Data Handling using Pandas in R

2.1 Refer Lab 2 Experiment G1. Repeat with R code.

2.2 List all the categorical (or nominal) attributes and the real-valued attributes separately.

i. For each attribute of Germen data set identify type of data and define data type, either numeric or string. a. If attribute is string type, find the values of attribute b. If the value is discrete, define attribute as nominal or categorical attribute. Otherwise, define attribute as string.

ii. Repeat step 1 until end of all attributes in data set.

iii. Display list of categorical and numerical valued attributes.

More exercises in data indexing and manipulation in R

Lab 5 Data description and manipulation using Data.Table

[http://r-tutorials.com/r-exercises-51-60-data-pre-processing-data-table/]

```
install.packages("data.table")
install.packages("ISLR") #for the dataset College

library(data.table)
library(ISLR)
```

I. 'College' dataset – Basic row manipulations

a. Transform 'College' from 'ISLR' to data.table. Make sure to keep the University identifier. We will use this new data.table called 'dtcollege' throughout this block of exercises.

b. Get familiar with the dataset and its variables.

c. Extract rows 40 to 60 as a new data.table ('mysubset').

d. What is the average enrollment number in this subset?

II. Advanced row selection

a. Get a data.table with all rows except the ones with an 'Outstate' fee between 8000-14000 USD.

III. Basic column operations

a. What are the top 10 Universities in terms of top 10% High School students (Top10perc)?

b. What are the top 10 Universities in terms of student enrollment vs. accepted applications (highest student enrollment ratio)? Add a new column to the data.table. Code this exercise step by step in several lines.

c. What are the top 10 Universities in terms of favorable S.F.Ratio with a Room.Board cost lower 4000 USD?

IV. Adding and removing new columns

a. Add a new column called 'HighInterest' to the data.table. The column has an integer 1 for all observations with a number of applications higher 1000.

b. Remove the 'HighInterest' column again.

V. Counting observations

a. How many Universities have instructional expenditures of over 20000 USD per year?

b. How many Universities have a combined 'Books' and 'Room.Board' costs of over 7000 USD per year?

c. How many Universities are public and how many are private?

VI. Working with keys and subsetting

a. Set two keys to your 'College' data.table: 'F.Undergrad' and 'P.Undergrad'. Check if the order has changed.

b. Get a subset of the 'College' data with 'F.Undergrad' lower 1000 and 'P.Undergrad' lower 100 students.

c. Is there a college with exactly 393 full-time and 4 part-time undergraduate students?

VII. Selecting existing columns and reshaping

a. Get a data.table with all columns except 'Apps', 'Accept', 'Enroll'. Use at least two different ways for this.

b. Get a data.table with the three columns 'Apps', 'Accept', 'Enroll'. Use at least two different code efficient methods.

VIII. Getting counts for grouped data

a. How many Colleges with less than 800 applications received, have a Top 10 student percentage above 40?

b. How many Colleges with less than 900 applications received and an 'Out of state tuition' below 10000, have a top 10 student percentage above 30?

c. How many Colleges with less than 1000 applications received, have a 'Top10perc' above 20 OR a 'Top25perc' above 30?

Lab 6 Basic programming and Data Manipulation in Python

Repeat the Exercises of Lab 3-4 in Python

Lab 7 Association Rule Mining – in R and Python

Experiment 1. Demonstration of Association rule process on two datasets using A. apriori algorithm. B. FP-Growth Algorithm.

Lab 8 Classification – Decision Tree – in R and Python

Q4. Demonstrate performance of classification on various data sets.
Experiment 9: Do you think it is a good idea to prefer simple decision trees instead of having long complex decision trees? How does the complexity of a Decision Tree relate to the bias of the model?

Experiment 10: You can make your Decision Trees simpler by pruning the nodes. One approach is to use Reduced Error Pruning - Explain this idea briefly. Try reduced error pruning for training your Decision Trees using cross-validation (you can do this in Weka) and report the Decision Tree you obtain? Also, report your accuracy using the pruned model. Does your accuracy increase?

[from weka]

Do you think it is a good idea to prefer simple decision trees instead of having long complex decision trees? How does the complexity of a Decision Tree relate to the bias of the model?

You can make your Decision Trees simpler by pruning the nodes. one approach is to use Reduced Error Pruning -Explain this idea briefly. Try reduced error pruning for training your Decision Trees using cross -validation (you can do this in Weka) and report the Decision Tree you obtain? Also, report your accuracy using the pruned model. Does your accuracy increase?

(Extra Credit): How can you convert a Decision Trees into "if –then -else rules". Make up your own small Decision Tree consisting of 2 - 3 levels and convert it into a set of rules. There also exist different classifiers that output the model in the form of rules -one such classifier in Weka is rules. PART, train this model and report the set of rules obtained. Sometimes just one attribute can be good enough in making the decision, yes, just one! Can you predict what attribute that might be in this dataset? OneR classifier uses a single attribute to make decisions (it chooses the attribute based on minimum error). Report the rule obtained by training a one R classifier. Rank the performance of j48, PART and oneR.

Lab 8 Classification – Bayesian Classification, SVM – in R/ Python

Experiment 1. Implement Bayesian Classification for various datasets
Experiment 2. Implement Classification using Support Vector Machines.

Lab 9 Clustering – in R/Python

Experiment 1. Demonstrate performance of clustering on two data sets using K-means clustering.
Experiment 2. Demonstrate performance of clustering on two data sets using Hierarchical clustering.
Experiment 3. Demonstrate performance of clustering on two data sets using SOM

Lab 10 Regression – in R/ Python

Experiment 1. Demonstrate performance of Regression on various data sets

Lab 11 Application of classification in Web mining – in R/Python

Experiment 1. Perform crowling using R and fetch features from the mined text
Experiment 2. Apply classification on the crowled web data

Lab 12 Data mining on Text data – Sentiment analysis – in R/Python

Experiment 1. Apply classification on the crowled web data

Lab 13 ETL process and Data Warehouse implementation

Q1. Build Data Warehouse/Data Mart (using open source tools like Pentaho Data Integration Tool, Pentaho Business Analytics; or other data warehouse tools like Microsoft-SSIS,Informatica,Business Objects,etc.,) A.(i) Identify source tables and populate sample data.

[https://help.hitachivantara.com/Documentation/Pentaho/9.5/Setup/Pentaho_Data_Integration_(PDI)_tutorial]

1. [https://mrcet.com/pdf/Lab%20Manuals/CSE%20IV-I%20SEM.pdf]

A.i. The data warehouse contains 4 tables: 1. Date dimension: contains every single date from 2006 to 2016. 2. Customer dimension: contains 100 customers. To be simple we'll make it type 1 so we don't create a new row for each change. 3. Van dimension: contains 20 vans. To be simple we'll make it type 1 so we don't create a new row for each change. 4. Hire fact table: contains 1000 hire transactions since 1st Jan 2011. It is a daily snapshot fact table so that every day we insert 1000 rows into this fact table. So over time we can track the changes of total bill, van charges, satnav income, etc.

A.(ii). Design multi-demesional data models namely Star, Snowflake and Fact Constellation schemas for any one enterprise (ex. Banking,Insurance, Finance, Healthcare, manufacturing, Automobiles,sales etc).

A.(iii) Write ETL scripts and implement using data warehouse tools.

A.(iv) Perform Various OLAP operations such slice, dice, roll up, drill up and pivot.

2. https://www.codeproject.com/Articles/652108/Create-First-Data-WareHouse

https://hevodata.com/learn/sql-server-for-data-warehouse-4-easy-steps/

https://docs.oracle.com/cd/B31080_01/doc/owb.102/b28223/concept_basics.htm

https://www.sircrrengg.ac.in/images/newsletter/ITMATERIALS/DMLab.pdf