TKR COLLEGE OF ENGINEERING AND TECHNOLOGY
(AUTONOMOUS)
(Sponsored by TKR Educational Society , Approved by AICTE, Affiliated by JNTUH,
Accredited by NBA & NAAC with 'A' Grade)

COMPUTER SCIENCE & ENGINEERING

B.Tech.VI Semester **L/T/P/C 0/0/2/1**

DATA WAREHOUSING AND DATA MINING LAB (C56PC5)

Course Objective:

Ability to understand the various kinds of tools, demonstrate the classification, clusters and etc. inlarge data sets.

Course Outcomes:

Upon completion of course, the student will be able to

1. Build a data warehouse and query it (using open source tools like PentahoData Integration and Pentaho Business Analytics)-L4
2. Understand data mining tasks using a data mining toolkit (such as open-sourceWEKA)-L2
3. Understand the data sets and data preprocessing-L2
4. Demonstrate the working of algorithms for data mining tasks such association rule mining, classification, clustering, and regression-L4
5. Practice the data mining techniques with varied input values for different parameters.- L4

TASK I:

Build Data Warehouse and Explore WEKA

Build a Data Warehouse/Data Mart (using open source tools like Pentaho Data Integration tool, Pentoaho Business Analytics; or other data warehouse tools like Microsoft-SSIS, Informatica, Business Objects etc.).

1. Identify source tables and populate sample data
2. Design multi-dimensional data models namely Star, snowflake and Fact constellation schemas for anyone enterprise (ex. Banking, Insurance, Finance, Healthcare, Manufacturing, Automobile, etc.).
3. Write ETL scripts and implement using data warehouse tools
4. Perform various OLAP operations such as slice, dice, roll up, drill up and pivot
5. Explore visualization features of the tool for analysis like identifying trends etc.

B. Explore WEKA Data Mining/Machine Learning Toolkit

1. Downloading and/or installation of WEKA data mining toolkit,
2. Understand the features of WEKA toolkit such as Explorer, Knowledge Flow interface, Experimenter, command-line interface.
3. Navigate the options available in the WEKA (ex. Select attributes panel, Preprocess panel, Classify panel, Cluster panel, Associate panel and Visualize panel) Study the arff file format

4. Explore the available data sets in WEKA.
5. Load a data set (ex. Weather dataset, Iris dataset,etc.)
6. Load each dataset and observe the following:
    i. List the attribute names and their types
    ii. Number of records in each dataset
    iii. Identify the class attribute (if any)
    iv. Plot Histogram
    v. Determine the number of records for each class.
    vi. Visualize the data in various dimensions

TASK II:

Perform data preprocessing tasks and Demonstrate performing association rule mining on data sets

1. Explore various options available in Weka for preprocessing data and apply (like Discretization Filters, Resample filter, etc.) on each dataset
2. Load each dataset into Weka and run Apriori algorithm with different support and confidence values. Study the rules generated.
3. Apply different discretization filters on numerical attributes and run the Apriori association rule algorithm. Study the rules generated. Derive interesting insights and observe the effect of discretization in the rule generation process.

TASK III:

Demonstrate performing classification on data sets

1. Load each dataset into Weka and run Id3, J48 classification algorithm. Study the classifier output. Compute entropy values, Kapp a statistic.
2. Extract if-then rules from the decision tree generated by the classifier, Observe the confusion matrix and derive Accuracy, F-measure, TPrate, FPrate, Precision and Recall values. Apply a cross-validation strategy with various fold levels and compare the accuracy results.
3. Load each dataset into Weka and perform Naïve-Bayes classification and k-Nearest Neighbour classification. Interpret the results obtained.
4. Plot RoC Curves
5. Compare classification results of ID3, J48, Naïve-Bayes and k-NN classifiers for each dataset, and deduce which classifier is performing best and poor for each dataset and justify.

TASK IV:

Demonstrate performing clustering on data sets

1. Load each dataset into Weka and run a simple k-means clustering algorithm with different values of k (number of desired clusters). Study the clusters formed. Observe the sum of squared errors and centroids, and derive insights.
2. Explore other clustering techniques available in Weka.
3. Explore visualization features of Weka to visualize the clusters. Derive interesting insights and explanations.

TASK V:

Demonstrate performing Regression on datasets

1. Load each dataset into Weka and build a Linear Regression model. Study the clusters formed. Use the Training set option. Interpret the regression model and derive patterns and conclusions from the regression results.
2. Use options cross-validation and percentage split and repeat running the Linear Regression Model. Observe the results and derive meaningful results.
3. Explore a Simple linear regression technique that only looks at one variable.

Resource Sites:

1. http://www.pentaho.com/
2. http://www.cs.waikato.ac.nz/ml/weka/

TASK VI:

Credit Risk Assessment Description:

The business of banks is making loans. Assessing the credit worthiness of an applicant is of crucial importance. You have to develop a system to help a loan officer decide whether the credit of a customer is good, or bad. A bank‘s business rules regarding loans must consider two opposing factors. On the one hand, a bank wants to make as many loans as possible. Interest on these loans is the banks profit source. On the other hand, a bank cannot afford to make too many bad loans. Too many bad loans could lead to the collapse of the bank. The bank‘s loan policy must involve a compromise: not too strict, and not too lenient. To do the assignment, you first and foremost need some knowledge about the world of credit. You can acquire such knowledge in a number of ways.

1. Knowledge Engineering. Find a loan officer who is willing to talk. Interview her and try to represent her knowledge in the form of production rules.
2. Books. Find some training manuals for loan officers or perhaps a suitable textbook on finance. Translate this knowledge from text form to production rule form.
3. Common sense. Imagine yourself as a loan officer and make up reasonable rules which can be used to judge the credit worthiness of a loan applicant.
4. Case histories. Find records of actual cases where competent loan officers correctly judged when, and when not to, approve a loan application.

The German Credit Data:

Actual historical credit data is not always easy to come by because of confidentiality rules. Here is one such dataset, consisting of 1000 actual cases collected in Germany. Credit dataset (original) Excel spreadsheet version of the German credit data. In spite of the fact that the data is German, you should probably make use of it for this assignment. (Unless you really can consult a real loan officer!)

A few notes on the German Dataset

1. DM stands for Deutsche Mark, the unit of currency, worth about 90 cents Canadian (but looks and acts like a quarter).
2. owns_telephone. German phone rates are much higher than in Canada so fewer people own telephones.
3. foreign_worker. There are millions of these in Germany (many from Turrkey). It is very hard to get German citizenship if you were not born of German parents. • There are 20 attributes used in judging a loan applicant. The goal is the classify the applicant into one of two categories, good or bad.

Subtasks: (Turn in your answers to the following tasks)

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.

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.
4. Suppose you use your above model trained on the complete dataset, and classify credit good/bad for each of the examples in the dataset. What % of examples can you classify

correctly? (This is also called testing on the training set) Why do you think you cannot get 100% training accuracy?

5. Is testing on the training set as you did above a good idea? Why or Why not ?

6. 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?

7. 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.

8. 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. (You had removed two attributes in problem 7. Remember to reload the arff data file to get all the attributes initially before you start selecting the ones you want.)

9. 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)?

10. 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?

11. 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?

12. (Extra Credit): How can you convert a Decision Trees into ―if-thenelse 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 one R.

Task Resources:

Mentor lecture on Decision Trees

Andrew Moore‘s Data Mining Tutorials (See tutorials on Decision Trees and Cross Validation)

Decision Trees (Source: Tan, MSU)
Tom Mitchell‘s book slides (See slides on Concept Learning and Decision Trees)

Weka resources:
Introduction to Weka (html version) (download ppt version) Download Weka Weka Tutorial ARFF format
Using Weka from command line

TASK VII:
Hospital Management System
Data Warehouse consists Dimension Table and Fact Table. REMEMBER The following Dimension The dimension object (Dimension):
_ Name_Attributes (Levels), with one primary key
_Hierarchies
One time dimension is must. About Levels and Hierarchies
Dimension objects (dimension) consist of a set of levels and a set of hierarchies defined over those levels. The levels represent levels of aggregation. Hierarchies describe parent-child relationships among a set of levels. For example, a typical calendar dimension could contain five levels. Two hierarchies can be defined on these levels:
H1: YearL > QuarterL > MonthL > WeekL > DayL H2: YearL > WeekL > DayL
The hierarchies are described from parent to child, so that Year is the parent of Quarter, Quarter the parent of Month, and so forth. About Unique Key Constraints When you create a definition for a hierarchy, Warehouse Builder creates an identifier key for each level of the hierarchy and a unique key constraint on the lowest level (Base Level) Design a Hospital Management system data warehouse (TARGET) consists of Dimensions Patient, Medicine, Supplier, Time. Where measures are _NO UNITS‘, UNIT PRICE. Assume the Relational database (SOURCE) table schemas as follows
TIME (day, month, year),
PATIENT (patient_name, Age, Address, etc.,)
MEDICINE (Medicine_Brand_name, Drug_name, Supplier, no_units,Uinit_Price,etc.,) SUPPLIER: (Supplier_name, Medicine_Brand_name, Address, etc.,) If each Dimension has 6 levels, decide the levels and hierarchies, Assume the level names suitably. Design the Hospital Management system data warehouse using all schemas. Give the example 4-D cube with assumption names.