CSCE 617

Take Home Exam (30%)

Deadline: Monday, Dec 14 @6PM
Contents

1. Problem Statement
2. Introduction
3. Tasks
4. Submission
5. Datasets
6. Resources
Problem Statement
Objective

To simulate big data benchmarks for modeling Classification Problem for Data Analytics
Introduction
Motivation

Big data is now a reality:

- The volume, variety and velocity of data coming into one organization and daily lives continue to reach unprecedented levels.
- This phenomenal growth means that not only must one understand big data in order to decipher the information that truly counts, but you also must understand the possibilities of big data analytics.

What is big data analytics?

Process of examining big data to uncover

- hidden patterns,
- unknown correlations and
- other useful information that can be used to make better decisions
Tasks
SubTask 1: Preprocessing + Data Generation + Pre-Visualization

Data Analytic Framework
SubTask 1: Preprocessing + Data Generation + Pre-Visualization

WHAT?

- To create a Big Dataset from a given real dataset that will be used later for training and validating predictive models.
- Synthetically created big data set should be such that it preserves the statistical properties of the given sample as accurately as possible, but without any data duplication.
SubTask 1

HOW?

● Use WEKA - Data Mining Software (GUI)
● Apply Clustering algorithm on the given dataset to find K clusters or groups.
● The entire data is distributed or categorized into these K clusters.
● If class labels are provided in dataset as an attribute, ignore that attribute for clustering. [Option available in GUI]
● Evaluate the computed clusters with respect to the class. Keep experimenting with parameters until incorrectly clustered instances are minimized.
● [OPTIONAL] You may use DBSCAN algorithm too if no attribute in the data corresponds to a class label.
● You may have to do some preprocessing of data before applying clustering. See section: Resource#3.
SubTask 1 continued

1. Each cluster is represented by its centroid i.e. values for the various attributes averaged over all data points belonging to that cluster.

2. Visualize the cluster using cluster visualization tool in WEKA and save the cluster assignments in a file. (default: arff format) “clusterAssign.arff”. Convert this arff format to csv format, “Real_train.csv”, where last column/attribute is for assigned cluster labels.

3. Write a program (Python) to generate box plot (See Resource#8) that displays underlying data for each attribute (y-axis) across cluster ids (x-axis). Save as “boxplot_<attributename>.png”

4. Analyze each of these box plots for different attributes and report each cluster’s characteristics.

5. Using the centroids obtained, write a program (C++ / Python/Java) to populate each cluster respectively to create synthetic data points. These new points should satisfy the statistical properties of their respective clusters. Each cluster should be populated to $3x$ of its original size.

6. The program must create and save the synthetic data along with their cluster ids in a csv file- “Synthetic_train.csv”. In addition, it should also create final training data “trainingData.csv” by merging data from both “Real_train.csv” and “Synthetic_train.csv”. Be careful that attributes ordering is consistent in all the three files for merge to take place.
Sub Task 2: Training + Validation + Post-Visualization
SubTask 2: Training + Validation + Post-Visualization

Data Analytic Framework
SubTask 2

WHAT?

- Create a Classifier Model and Validate it until good accuracy is obtained.
- Objective is to predict class labels (the ones obtained by Clustering in subtask 1).
- Compare performance (accuracy in prediction) of 3 classifiers
  - Naive Bayes Classifier
  - Support Vector Machines [LibSVM in Weka]
  - Random Forest (type of ensemble learning where weak classifiers are grouped together and decision/prediction is made on majority vote)
SubTask 2: Details into Modeling

Classification Step:
Train, validation, and test

You are Here
[trainingData.csv]
SubTask 2

HOW?

- Convert "trainingData.csv" to corresponding arff format (to be used in WEKA).
- This data has to be divided into two sub-datasets: one part to train the model and other to validate it (as shown in diagram in previous slide). Generally, Training Set: 70-80% and Validation Set: 20-30%. [Use GUI in WEKA to do it].

[Behind the Scene] Model is getting trained on the Training set by learning the attributes and corresponding class labels. Once it is trained, it is used to predict the class labels of the records in the Validation Set. These predicted class labels are evaluated against the actual class labels of these records. The goal is to keep experimenting (or modeling the parameters) until the validation error obtained is minimized.

- Use 3 classifier models and compare their performance (error rate, accuracy).
- To train a model accurately, keep experimenting with the classifier parameters to get low training and validation errors, thereby improving overall accuracy.
Submission

1. All report, files and codes (see next page) need to be turned in as a zip file `CSCE617_GroupID.zip` to rabi@cse.tamu.edu and cc to jyoti1991@tamu.edu

2. **Deadline:** Monday, Dec 14 @6PM

3. For any queries related to tasks, contact Jyotikrishna Dass “JD” jyoti1991@tamu.edu
Deliverables

● SubTask 1 [15%]
  1) Report parameters of trained K-Means clustering model: K, error, iterations, cluster centroids, distribution percentage of data into clusters
  2) `Real_train.csv`, `Synthetic_train.csv`, `trainingData.csv`
  3) Submit the code for box plot generation. Also, include plots `boxplot_<attributename>.png` for every attribute used in clustering. Analyze these plots and report each cluster’s characteristics based on the attributes.
  4) Submit the statistical properties (mean, standard deviation, min, max, median) of the real and synthetic data.
  5) Submit the code for synthetic data generation.
  6) Report the data dimensions of Real and Synthetic datasets (#records, #features).

● SubTask 2 [3 classifiers x 5% = 15%]
  For each classifier model = {Naive Bayes, LibSVM, Random Forest}, report
  a) Training Data and Validation data percentage used
  b) Briefly explain the model parameters (1-2 lines) and report the values used.
  c) In a table, compare the accuracy of prediction, Training error, Validation (or prediction) error
  d) Describe the class attribute you are predicting through this model.
  e) Submit any plots for performance analysis (eg: ROC, confusion matrix etc)
Datasets

- 7 Groups, each with 2 students
- Group# 1, 4, 7: **DataSet1**
- Group# 2, 5: **DataSet2**
- Group# 3, 6: **DataSet3**

Group 1: Richa Surbhi + Rohit
Group 2: Ghanshyam + Siddhant + Saket
Group 3: Ranjita + Shubhankar
Group 4: Ankit + Karl
Group 5: Divyesh + Odair
Group 6: Richa Deo + Pritam
Group 7: Luna + Vighnesh
Resources

Follow tutorials - step by step guides to get acquainted with WEKA interface

6. Tutorials and Video Lectures
7. LibSVM: [https://weka.wikispaces.com/LibSVM#Examples](https://weka.wikispaces.com/LibSVM#Examples)