Data Mining Project 3 Report

Prepared by:

Yashodhan Kumthekar 1001544391

Pradnya Gaikwad 1001626964

Course:

2182-CSE-5334-001-DATA-MINING--2018-Spring

Professor:

Chris H.Q. Ding, Ph.D.
Department of Computer Science and Engineering
University of Texas at Arlington

Teaching Assistant:

Qicheng Wang

Date:

9th May 2018

Tasks to be performed:

- Feature selection using F-Test method
- Using the top 100 features from the f-test to classify the test data using
    - SVM
    - Linear Regression
    - K-Nearest Neighbor
    - Centroid Algorithm

DataSets:

1. Training Data:

   1st row is Labels

   2nd- to end row: Sample data

   Total 40 data instances and 4434 features

2. Test Data:

   Total 10 data instances. With 4434 features.

Task A.

Use GenomeTrainXY.txt to select 100 top-ranked genes based on f-test. You need to submit the f-test scores and the feature number (the line/feature number).

Task B.

Use the above selected genes as the features, train the four classifiers
a: SVM linear kernel
b: linear regression
c: KNN (k=3)
d: centroid method

Task A.

Use trained classifiers to predict the class labels of data instances provided in GenomeTestX.txt

Observations:

1. We were able to predict classes of un-labeled data using a classifier trained using labelled data.

Results:

1. Results are provided in results.txt file.