Team Project Final Report

11.29.2019

Cyrus Baker, Lucas Herrmann, Ivan Nieto
CS 488 - Data Mining

Data

Our data is based on breast cancer tumor statistics collected from the University of Wisconsin hospitals. The data contains 699 rows from 699 patients, and includes tumor characteristics from clump thickness to mitoses. The data was collected from January 1989 to November 1991.

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

Data Mining Task

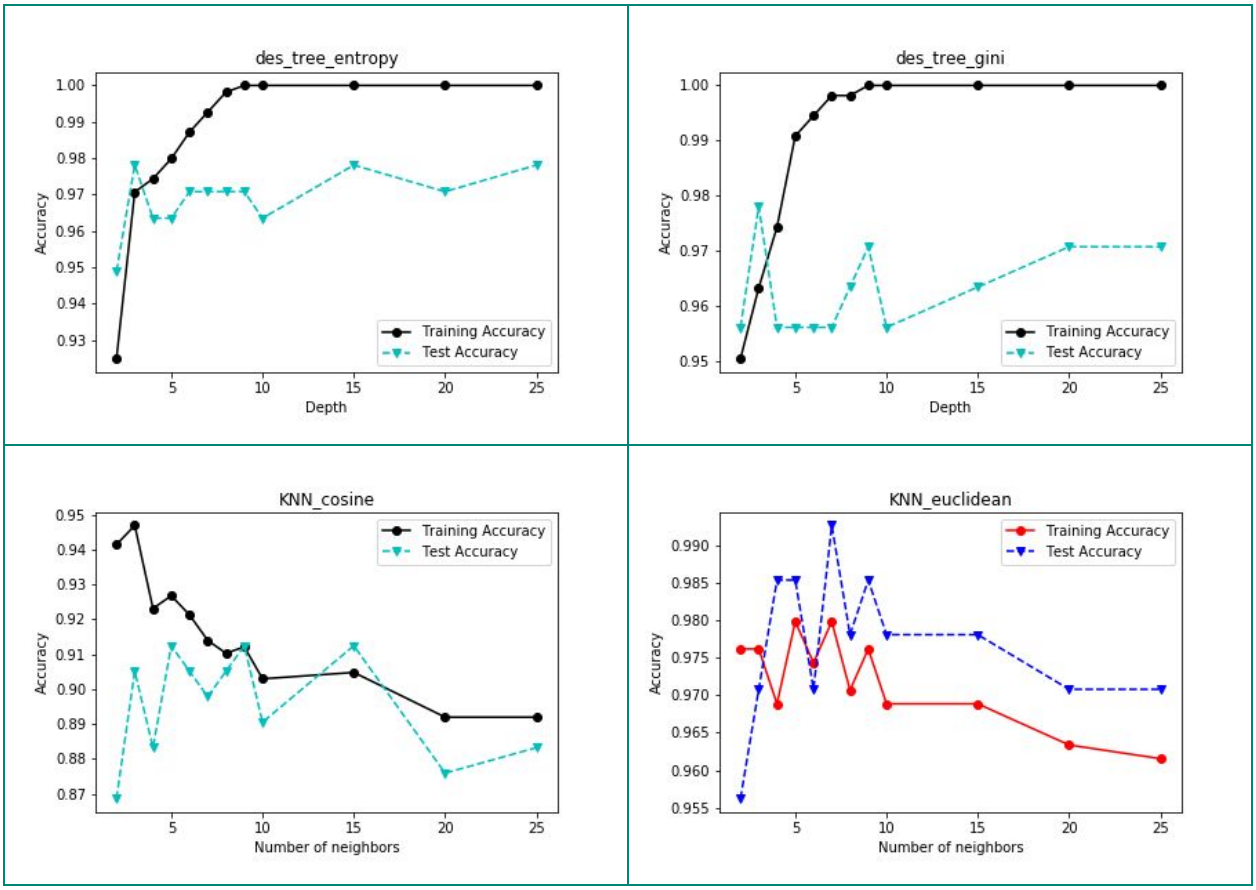
We are performing a classification tasks on our breast cancer data set. We use the first few attributes (clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses) to predict the classification of the tumor. To achieve this classification, we have chosen to train and test the data using a decision tree classifier and a KNN classifier.

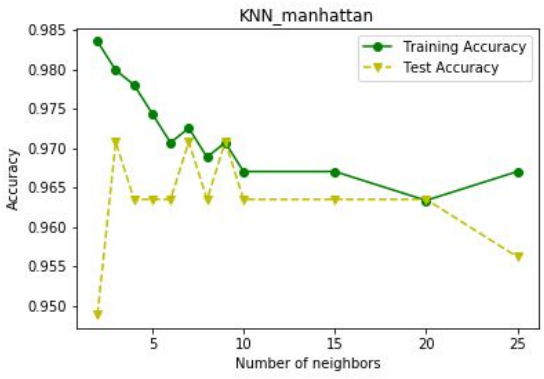
The data from the study has the following instances:

```
Data columns (total 11 columns):
    Sample Code #                   699 non-null int64
    Clump Thickness                 699 non-null int64
    Uniformity of Cell Size         699 non-null int64
    Uniformity of Cell Shape        699 non-null int64
    Marginal Adhesion               699 non-null int64
    Single Epithelial Cell Size     699 non-null int64
    Bare Nuclei                     699 non-null object
    Bland Chromatin                 699 non-null int64
    Normal Nucleoli                 699 non-null int64
    Mitoses                         699 non-null int64
    Class - 2=Benign 4=Malignant    699 non-null int64
```

Implementation and Experimental Results

To date, we first cleaned the data by removing the ID attribute from the data set. After this we removed any null values from the dataset that were represented by a '?' character. We then split the data into test and training data, and constructed both a KNN classifier and a decision tree classifier. To achieve this, we used the pandas and numpy python libraries to process the data, and the sklearn library to construct our classifiers much like we did in our second homework.





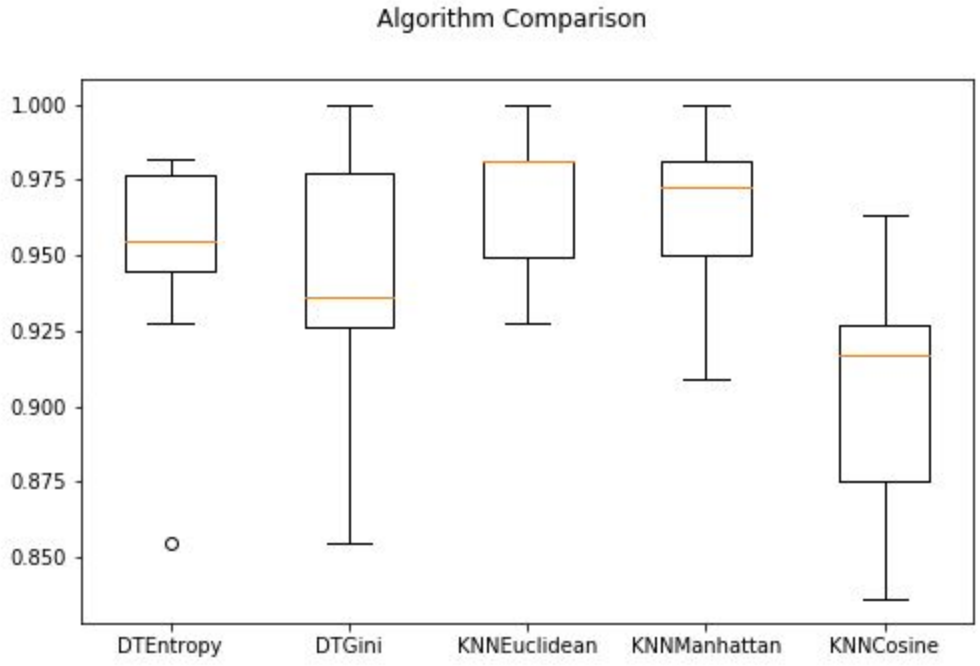
- 10 Fold cross validation scores in order:

1. Decision tree Entropy (94.89%)
2. Decision tree Gini (93.59%)
3. K-Nearest Neighbor Manhattan (96.53%)
4. K-Nearest Neighbor Euclidean (96.71%)
5. K-Nearest Neighbor Cosine (90.3%)

Best case was K-Nearest Neighbor classifier with Euclidean Distance and 5 neighbors

We had a slightly higher cross validation score when we applied IQR outlier removal to the dataset but that came with a significant decrease in the accuracy score of the classifiers so we decided to stick with z score outlier detection.

Running times between algorithms were not substantially different enough for it to factor into our decision of choosing the K-Nearest Neighbor classifier with Euclidean Distance as our best implementation.


Algorithm Comparison
1.000
0.975
0.950
0.925
0.900
0.875
0.850
DTEntropy
DTGini
KNNEuclidean
KNNManhattan
KNNCosine