PROJECT FINAL PHASE
CSE 572
DATA MINING
SPRING 2018

SUBMITTED TO:

Professor Ayan Banerjee
Ira A. Fulton School of Engineering
Arizona State University

A Report by:
Athithyaa Selvam (aselvam@asu.edu)
Hari Siddarth Velaudampalayam Kesavan (hvelauda@asu.edu)
Mohan Vasantrao Yadav (mvasantr@asu.edu)
Raam Prashanth Namakkal Sudhakar (rnamakka@asu.edu)
Sangeetha Swaminathan (sswami11@asu.edu)
(Team 18)

TABLE OF CONTENTS

1. DATA PREPARATION:

We worked on data collected from multiple users. The feature matrix from phase 2 is used as input for the three classification techniques. Training is performed together on data collected from 10 users and testing is done on multiple users.
Number of Features used: 10

2. SUPPORT VECTOR MACHINES (SVM)

Support Vector Machine is a supervised learning model used for classifying a data point to one of the two available classes. A training dataset with two different classes is used to build the SVM model and this model is used to further classify the given data point to one of the two classes accurately. The algorithm generates an optimal hyperplane separating the two classes from the training data.

File: Code/task4svm.m

2.1. Steps to classify data using SVM

1. 10 user data is used for training and rest of the data is used for testing.
2. SVM model is built based on the training data and training labels using the matlab inbuilt function fitcsvm().
3. The inbuilt predict() function is called which takes in the SVM model constructed and the test data and it returns the predicted label.
4. The predicted labels are compared with actual test labels using the inbuilt function confusionmat() which generates the confusion matrix.
5. Values from the confusion matrix are used to calculate the precision, recall and F-1 Score.

2.2. Accuracy Metrics for each user using SVM

Table 2.2.1 - SVM Accuracy metrics for user 1

Class	Precision	Recall	F-1 Score
About	0.9615384615	0.641025641	0.7692307692
And	0.8695652174	0.6666666667	0.7547169811
Can	0.9090909091	0.7142857143	0.8
Cop	0.6363636364	0.7368421053	0.6829268293
Deaf	0.4285714286	0.6923076923	0.5294117647
Decide	0.9523809524	0.7692307692	0.8510638298
Father	0.95	0.7916666667	0.8636363636
Find	0.9090909091	0.7407407407	0.8163265306
Go Out	0.8345	0.7483	0.782
Hearing	0.5238095238	0.6875	0.5945945946

Table 2.2.2 - SVM Accuracy metrics for user 2

Class	Precision	Recall	F-1 Score
About	1	0.5777777778	0.7323943662
And	0.8218	0.7208	0.791
Can	1	0.6285714286	0.7719298246
Cop	0.1363636364	0.2727272727	0.1818181818
Deaf	0.5714285714	0.6666666667	0.6153846154
Decide	0.9523809524	0.6896551724	0.8
Father	0.1	0.25	0.1428571429
Find	0.1363636364	0.1875	0.1578947368
Go Out	0.923	0.854	0.89
Hearing	0.4761904762	0.5263157895	0.5

Table 2.2.3 - SVM Accuracy metrics for user 3

Class	Precision	Recall	F-1 Score
About	0.9615384615	0.6944444444	0.8064516129
And	0.923	0.854	0.89
Can	1	0.7857142857	0.88
Cop	0.9090909091	0.7407407407	0.8163265306
Deaf	0.4285714286	0.6428571429	0.5142857143
Decide	0.9523809524	0.7692307692	0.8510638298
Father	0.85	0.8095238095	0.8292682927
Find	0.8181818182	0.75	0.7826086957
Go Out	0.65	0.565	0.604
Hearing	0.4761904762	0.625	0.5405405405

3. DECISION TREE

Decision tree is a greedy classification approach in which a tree like model is developed where each node is split based on a condition and the leaves represent the possible outcome classes of that model. Using this model the user can understand the target classes just by observation.

File Name: Code/task4dt.m

3.1. Steps to classify data using Decision Tree

1. 10 user data is used for training and rest of the data is used for testing.
2. Decision Tree is built based on the training data and training labels using the matlab inbuilt function fitctree().
3. The inbuilt predict() function is called which takes in the tree generated and the test data and it returns the predicted label.
4. The predicted labels are compared with actual test labels using the inbuilt function confusionmat() which generates the confusion matrix.
5. Values from the confusion matrix are used to calculate the precision, recall and F-1 Score.

3.2. Accuracy Metrics for each user using Decision Tree

Table 3.2.1 - Decision tree accuracy metrics for user 1

Class	Precision	Recall	F-1 Score
About	1	0.5652173913	0.7222222222
And	0.9565217391	0.5	0.6567164179
Can	0.5909090909	0.7647058824	0.6666666667
Cop	0.8636363636	0.5428571429	0.6666666667
Deaf	0.9047619048	0.7307692308	0.8085106383
Decide	0.9523809524	0.6451612903	0.7692307692
Father	1	0.7407407407	0.8510638298
Find	0.9545454545	0.6	0.7368421053
Go Out	0.4	1	0.5714285714
Hearing	1	0.5675675676	0.724137931

Table 3.2.2 - Decision tree accuracy metrics for user 2

Class	Precision	Recall	F-1 Score
About	0.9615384615	0.641025641	0.7692307692
And	0.9130434783	0.488372093	0.6363636364
Can	0.7272727273	0.8	0.7619047619
Cop	0.8636363636	0.6551724138	0.7450980392
Deaf	0.8571428571	0.6428571429	0.7346938776
Decide	1	0.65625	0.7924528302
Father	0.65	0.5652173913	0.6046511628
Find	1	0.6111111111	0.7586206897
Go Out	0.65	0.574	0.6046511628
Hearing	0.9523809524	0.6060606061	0.7407407407

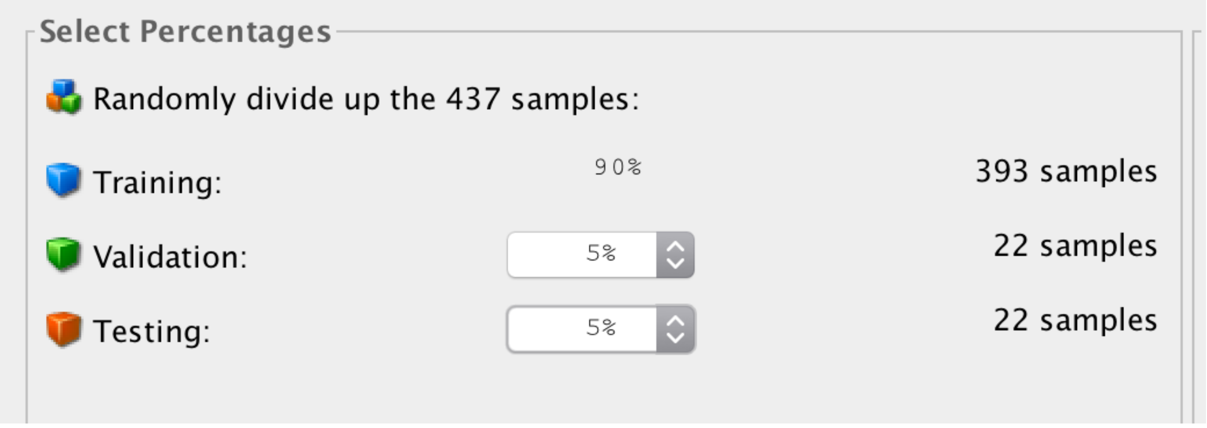
Table 3.2.3 - Decision tree accuracy metrics for user 3

Class	Precision	Recall	F-1 Score
About	0.8846153846	0.6764705882	0.7666666667
And	0.3043478261	0.2916666667	0.2978723404
Can	0.7727272727	0.7727272727	0.7727272727
Cop	0.3636363636	0.4705882353	0.4102564103
Deaf	0.5238095238	0.6111111111	0.5641025641
Decide	0.4285714286	0.45	0.4390243902
Father	0.6	0.6666666667	0.6315789474
Find	0.04545454545	0.09090909091	0.06060606061
Go Out	0.8	0.6666666667	0.7272727273
Hearing	0.5714285714	0.5714285714	0.5714285714

4. NEURAL NETWORK

Neural Networks are used to find patterns in data by creating a complex model relationship between inputs and outputs. Neural Networks are prominent for being versatile, i.e., they change themselves as they progress from starting, preparing and ensuing runs to give more data about the world.

File: Code/task4nn.m



Select Percentages

Randomly divide up the 437 samples:

Training: 90% 393 samples

Validation: 5% 22 samples

Testing: 5% 22 samples

Fig 4.1 - Splitting of data

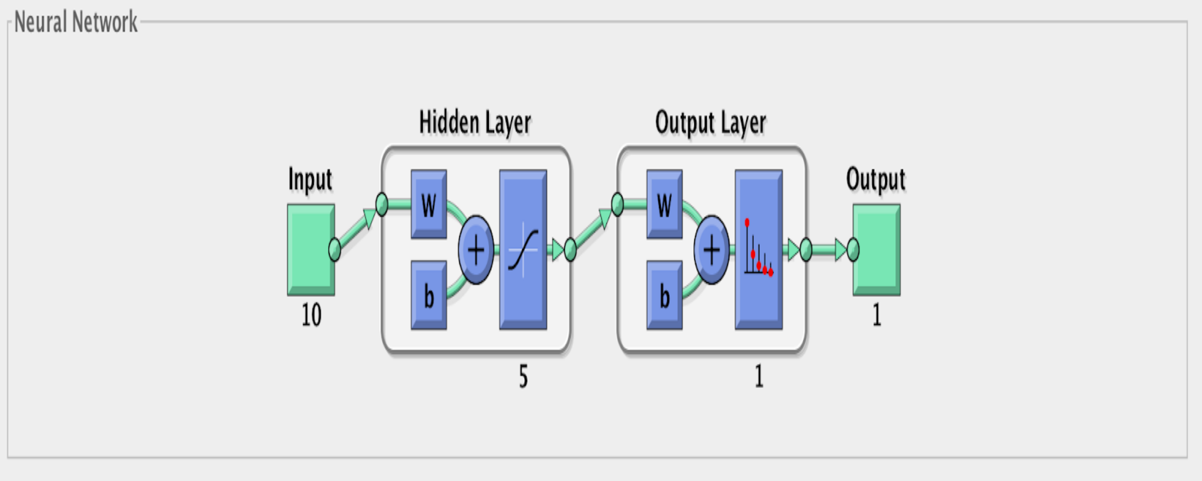


Fig 4.2 - Constructed neural network

4.1. Steps to classify data using Neural Network:

1. The data is divided into training, test and validation based on the percentage indicated as mentioned in Fig 4.1 using Neural Network Toolbox.
2. Neural Network is trained using scaled conjugate gradient back-propagation.
3. The Confusion matrix is plotted and the Precision, Recall and F-1 Scores are calculated.

4.2. Accuracy Metrics for each user using Neural Network:

Table 4.2.1 - Neural network accuracy metrics for user 1

Class	Precision	Recall	F-1 Score
About	0.875	0.875	0.875
And	0.9090909091	0.9090909091	0.9090909091
Can	0.9	0.8181818182	0.8571428571
Cop	0.8181818182	0.8181818182	0.8181818182
Deaf	0.9285714286	0.8666666667	0.8965517241
Decide	0.9166666667	0.9166666667	0.9166666667
Father	0.6923076923	0.9	0.7826086957
Find	0.8888888889	0.8	0.8421052632
Go Out	0.6666666667	0.6666666667	0.6666666667
Hearing	0.9375	0.9375	0.9375

Table 4.2.2 - Neural network accuracy metrics for user 2

Class	Precision	Recall	F-1 Score
About	0.9285714286	0.9285714286	0.9285714286
And	0.6666666667	0.8571428571	0.75
Can	0.9230769231	0.75	0.8275862069
Cop	0.8461538462	0.7857142857	0.8148148148
Deaf	0.8181818182	0.6923076923	0.75
Decide	0.8181818182	0.8181818182	0.8181818182
Father	0.6363636364	0.7777777778	0.7
Find	0.7142857143	0.7692307692	0.7407407407
Go Out	0.5	0.5	0.5
Hearing	0.875	0.7777777778	0.8235294118

Table 4.2.3 - Neural network accuracy metrics for user 3

Class	Precision	Recall	F-1 Score
About	0.7647058824	0.9285714286	0.8387096774
And	0.7777777778	0.875	0.8235294118
Can	0.9166666667	0.9166666667	0.9166666667
Cop	0.8181818182	0.75	0.7826086957
Deaf	0.9166666667	0.9166666667	0.9166666667
Decide	0.7777777778	0.9333333333	0.8484848485
Father	0.6666666667	0.75	0.7058823529
Find	0.8888888889	0.6666666667	0.7619047619
Go Out	1	1	1
Hearing	0.9285714286	0.8666666667	0.8965517241

5. CONCLUSION

From our observation of the results produced from the three techniques, User independent analyses also produces better accuracy with Neural Network classification compared to Support Vector Machine and Decision Tree.