Group 24

Project 4 Report

Data Mining

CSE 572 Spring 2018

Submitted to:

Professor Ayan Banerjee

Ira A. Fulton School of Engineering

Arizona State University

Submitted by:

Anuhya Sai Nudurupati (anuduru@asu.edu)

Anvitha Dinesh Rao (arao30@asu.edu)

Rachana Kashyap (rnkashya@asu.edu**)**

Rohit Balachandran Menon (skolli6@asu.edu**)**

Vaishak Ramesh Vellore (vvellore@asu.edu**)**

1 Introduction

In this project, we attempt to develop a system which can understand and recognize the American Sign Language(ASL) through human gestures. A wristband sensor worn on both hands is used to collect data related to acceleration, gyroscope, orientation, electromyography and kinect data and is mined to understand what gesture the person has made. This could help a person who does not understand ASL to be able to communicate with a deaf/dumb person who does communicate in ASL. We use MATLAB to develop this software.

2 Project Phase 1

In the first phase, we went to the IMPACT lab at Brickyard, Tempe in order to collect data. One person wore wrist bands on both arms and made the gestures, “ABOUT”, “AND”, “CAN”, “COP” ,“DEAF”,“DECIDE”, “FATHER”, “FIND”, “GO OUT” and “HEARING” about 20 times each. The data collected from the sensors is stored in the form of CSV files. The time series data is sampled every 3 seconds. The frequency of sensors was found to be 15Hz. The data headers of the collected data are Accelerometer, Electromyogram, Gyroscope and Orientation.

3 Project Phase 2

The second phase of the project involves feature extraction and feature selection aspects of Data Mining. PCA was applied to the feature matrix to obtain the new feature matrix . From the feature matrix 7 features were extracted and multiplied with the feature matrix obtain a projection matrix . This projection matrix is used as a new feature matrix.

4 Project Phase 3

The third phase of the project involves the following steps,

A. A new column is added to the new feature matrix obtained from phase 2 for each user in order to create labels used for binary classification.

B. The data is shuffled and selected at random from the new feature matrix with labels generated.

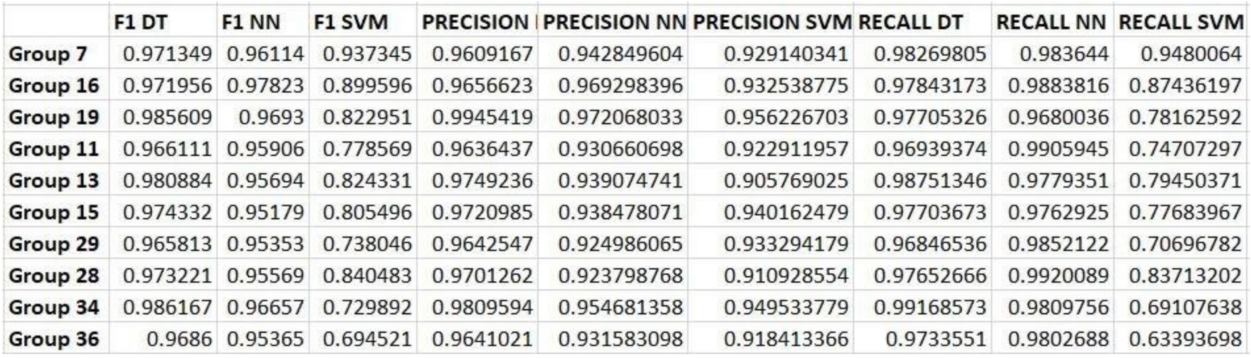
C. 60% of the data for each user was used for training.

D. 40% of the remaining data was used for testing.

E. Support Vector Machines, Neural Networks and Decision Trees were used for training the machine .

F. The test dataset is then used to obtain the accuracy metrics Precision, Recall and F1 score for each user.

Comparison Table of the average accuracy metrics for Decision Tree, Neural Networks and Support Vector Machine for 10 Users



	F1 DT	F1 NN	F1 SVM	PRECISION	PRECISION NN	PRECISION SVM	RECALL DT	RECALL NN	RECALL SVM
Group 7	0.971349	0.96114	0.937345	0.9609167	0.942849604	0.929140341	0.98269805	0.983644	0.9480064
Group 16	0.971956	0.97823	0.899596	0.9656623	0.969298396	0.932538775	0.97843173	0.9883816	0.87436197
Group 19	0.985609	0.9693	0.822951	0.9945419	0.972068033	0.956226703	0.97705326	0.9680036	0.78162592
Group 11	0.966111	0.95906	0.778569	0.9636437	0.930660698	0.922911957	0.96939374	0.9905945	0.74707297
Group 13	0.980884	0.95694	0.824331	0.9749236	0.939074741	0.905769025	0.98751346	0.9779351	0.79450371
Group 15	0.974332	0.95179	0.805496	0.9720985	0.938478071	0.940162479	0.97703673	0.9762925	0.77683967
Group 29	0.965813	0.95353	0.738046	0.9642547	0.924986065	0.933294179	0.96846536	0.9852122	0.70696782
Group 28	0.973221	0.95569	0.840483	0.9701262	0.923798768	0.910928554	0.97652666	0.9920089	0.83713202
Group 34	0.986167	0.96657	0.729892	0.9809594	0.954681358	0.949533779	0.99168573	0.9809756	0.69107638
Group 36	0.9686	0.95365	0.694521	0.9641021	0.931583098	0.918413366	0.9733551	0.9802688	0.63393698

5 Project Phase 4

The fourth phase of the project involves the following steps,

A. A new column is added to the new feature matrix obtained from phase 2 for all users not used in training and testing datasets in order to create labels used for binary classification.

B. The training data is shuffled and selected at random from the new feature matrix with labels generated.

C. The testing data is also shuffled and selected at random from the new feature matrix with labels generated.

D. 10 users data is used for training.

E. The rest of the user data is used for testing.

F. Support Vector Machines, Neural Networks and Decision Trees were used for training the machine .

G. The test dataset is then used to obtain the accuracy metrics Precision, Recall and F1 score for all users in the test data.

Code:

```
files = dir('*.csv');
i=1;
decision_tree_accuracy = zeros(1,10);
svm_res_accuracy = zeros(1,10);

for file=files'
csv=readtable(file.name,"ReadRowNames",false);

%%%%%%%%%%%%%% Pre-processing %%%%%%%%%%%%%%%

random_dataset = csv(randperm(size(csv, 1)), :);
split = floor(size(random_dataset,1)/3);
var=split*2+1;
train_data = random_dataset(1:split*2,:);
test_data = random_dataset(var:end,1:end);
test_set = test_data(:,1:end-1);
test_labels = test_data(:,end:end);
train_set = train_data(:,1:end-1);
train_labels = train_data(:,end:end);

%%%%%%%%%%%% DECISION TREE%%%%%%%%%%%%%%%%%

dt = fitctree(train_set,train_labels);
p = predict(dt, test_set);
decision_tree_accuracy(i) =
sum(table2array(test_labels)==p)/size(table2array(test_labels),1
;

[confMat1,order1] = confusionmat(table2array(test_labels), p);
recall_dt(i)=confMat1(1,1)/sum(confMat1(1,:));
precision_dt(i) = confMat1(1,1)/sum(confMat1(:,1));
f1_score_dt(i) =
2*recall_dt(i)*precision_dt(i)/(recall_dt(i)+precision_dt(i));
```

```
%%%%%%%%%%%%%%%%%%% SVM %%%%%%%%%%%%%%%%%%%%%%%

svm = fitcecoc(train_set,train_labels);
v = predict(svm, test_set);
svm_res_accuracy(i) =
sum(table2array(test_labels)==v)/size(test_labels,1);

[confMat2,order2] = confusionmat(table2array(test_labels), v);
recall_svm(i)=confMat2(1,1)/sum(confMat2(1,:));
precision_svm(i) = confMat2(1,1)/sum(confMat2(:,1));
f1_score_svm(i) =
2*recall_svm(i)*precision_svm(i)/(recall_svm(i)+precision_svm(i)
;

%%%%%%%%%%%%%%%%%%%%% NEURAL NETS %%%%%%%%%%%%%
net = patternnet(5);
net = train(net,table2array(train_set)', table2array(train_labels)');
y = round(net(table2array(test_set)'));
[confMat3,order3] = confusionmat(table2array(test_labels), y');
recall_nn(i)=confMat3(1,1)/sum(confMat3(1,:));
precision_nn(i) = confMat3(1,1)/sum(confMat3(:,1));
f1_score_nn(i) =
2*recall_nn(i)*precision_nn(i)/(recall_nn(i)+precision_nn(i));

%%%%%%%%%%%%%%NEXT FILE ITERATION%%%%%%%%%%%%%%%
i= i+1;
end

%%%%%%%%%%%%AVERAGE R and P DT %%%%%%%%%%%%%%%%%

Average_Recall_User_DT = mean(recall_dt);
Average_Precision_User_DT = mean(precision_dt);

%%%%%%%%% AVERAGE R and P SVM %%%%%%%%%%%%%%%%%%

Average_Recall_User_SVM = mean(recall_svm);
Average_Precision_User_SVM = mean(precision_svm);
%%%%%%%%%%%%%%AVERAGE R and P NN %%%%%%%%%%%%%%%%
Average_Recall_User_NN = mean(recall_nn);
Average_Precision_User_NN = mean(precision_nn);

%%%%%%%%%%%% AVERAGE F1 DT and SVM and NN %%%%%%%
Average_F1_DT = mean(f1_score_dt);
```

```
Average_F1_SVM = mean(f1_score_svm);
Average_F1_NN = mean(f1_score_nn);
```

Average accuracy metrics for Decision Tree, Neural-Networks and Support Vector Machine for remaining 27 Users

Average_F1_DT	0.927453172871062
Average_F1_NN	0.908953257553748
Average_F1_SVM	0.906247090999679
Average_Precision_DT	0.900000000000000
Average_Precision_NN	0.896608359930718
Average_Precision_SVM	0.900000000000000
Average_Recall_DT	0.956662602184637
Average_Recall_NN	0.923615504515750
Average_Recall_SVM	0.912608451832271

5. Conclusion:

Since there was a lot of noise in the data, such user data had to be discarded. With 10 good user directories used for the training the model, it is found that SVM performs much better here than in the user dependent analysis. It was found that the average accuracy metrics for user independent analysis is slightly lower than that of user dependent analysis.