

## Analysis for problem 2: Touch biometrics

\*\*\*For our analysis we have decided to neglect all rows with missing feature values. This was done to make analysis easier, better performance could probably be obtained by filling in these values using imputation with mean or regression on other features.

Apart from the 30 features suggested in the paper, we implemented two more features by adding a few more lines of code to the feature extraction matlab script. All the other analysis was performed with a python script. The new features are:

1.) Beginning to mid stroke pressure variation: This feature tries to measure the variation in pressure from the start to the middle of the stroke. We expect this to be different for distinct users since different people will tend to apply pressure in different ways (similar to how velocity varies over different users).

2.) Median Area Variation: This feature measures the variations in the area covered by the finger over the course of the entire stroke. This feature tries to capture if a user varies the finger position mid-stroke often.

### MUTUAL INFORMATION:

We implemented a function called `print_rel_entropy()` that calculates and prints the mutual information of all features with the user ids.

Our method differs slightly from the one in the paper in that instead of not accounting for the outliers (outside the 10 - 90%ile range), we place them in buckets closest to their values. We also neglect rows with missing features. As a result our observed mutual information values are slightly different from those observed in the paper.

The observed mutual information values are:

```
inter-stroke time 0.0515011140459
stroke duration 0.0811066790426
start $x$ 0.101068996739
start $y$ 0.0880174672413
stop $x$ 0.100919891785
stop $y$ 0.0783172515923
direct end-to-end distance 0.0697476333063
  mean resultant length 0.0495390012763
up/down/left/right flag 0.0156600490404
direction of end-to-end line 0.132147002313
20%-perc. pairwise velocity 0.0587495155367
50%-perc. pairwise velocity 0.0776533031399
80%-perc. pairwise velocity 0.0675999022843
20%-perc. pairwise acc 0.0436980814235
50%-perc. pairwise acc 0.0449835793715
80%-perc. pairwise acc 0.0392164148143
median velocity at last 3 pts 0.0581688573246
largest deviation from end-to-end line 0.0476323365402
20%-perc. dev. from end-to-end line 0.0441975732893
50%-perc. dev. from end-to-end line 0.0406942418574
80%-perc. dev. from end-to-end line 0.0388805467283
average direction 0.104251020783
length of trajectory 0.0694785739857
ratio end-to-end dist and length of trajectory 0.0470056765146
average velocity 0.0835889944846
```

median acceleration at first 5 points 0.0385777697036  
mid-stroke pressure 0.240112069392  
mid-stroke area covered 0.233210843741  
mid-stroke finger orientation 0.0252410461342  
phone orientation 0.0283888714367  
beginning to mid stroke pressure variation 0.176455732727  
median area variation 0.1883318454

As we see, the top 5 features with maximum mutual information with the user id are:

mid-stroke pressure 0.240112069392  
mid-stroke area covered 0.233210843741  
median area variation 0.1883318454 (new feature)  
beginning to mid stroke pressure variation 0.176455732727 (new feature)  
direction of end-to-end line 0.132147002313

As we can see, our new features have very high mutual information indicating that they are highly correlated with the user ids.

### **CORRELATION:**

We implemented a function `correlation()` that calculates and prints the correlation of all the features. The correlation of our new features with the other features are (features occur in the same order as in the mutual information table):

```
beginning to mid stroke pressure variation
[-0.004  0.221 -0.043 -0.091 -0.087  0.132  0.035 -0.094  0.116  0.113
 -0.227 -0.231 -0.141  0.048 -0.029 -0.094 -0.171 -0.032 -0.061 -0.05
 -0.022  0.089  0.064 -0.134 -0.2   -0.025  0.581 -0.223 -0.001  0.005  1.
 -0.341]
median area variation
[ 0.007 -0.152  0.008  0.036 -0.007 -0.106  0.043  0.032 -0.029 -0.113
 0.267  0.278  0.189 -0.079  0.054  0.129  0.209  0.015 -0.028  0.014
 0.052 -0.092  0.011  0.084  0.215  0.052 -0.18  0.584 -0.035  0.036
 -0.341  1.   ]
```

As we can see, our new features have very little correlation ( $<0.15$ ) with most of the features. However, they are highly negatively correlated with each other ( $-0.341$ ). Apart from that, beginning to mid-stroke pressure variation is highly correlated with mid stroke pressure ( $0.581$ ) and median area variation is highly correlated with mid stroke area covered ( $0.584$ ) which indicates that although they might be good predictors of user-id individually, in the presence of other correlated features they might not add as much. They also make sense since a area/pressure variations for a person with high area covered/pressure tend to be larger in magnitude. Using this variations relative to the original values could have been another useful signal.

### **CLASSIFICATION:**

We used 4 feature selection methods and tested the performance of the selected features ( + our features) with two classifiers – SVM with RBF kernel and a plain logistic regression classifier. To improve the performance of our classifiers, we normalized the values of all features to 0 mean and 1 Std. Deviation. We also trained different classifiers for each user and used all the other users as negative examples (or masqueraders). To balance the negative and positive examples we multiplied the weight of the features corresponding to the actual user by 40 (another way would have been to sample 2.5% of the masquerader examples but that would have resulted in loss of data). We list their performance for the 10 (+2) features and over the entire set of features below:

1.) The first selection method was highest mutual information. In this method we

selected the features with highest mutual information with the user-id without any regards to correlation with each other. The selected features and their performance were:

mid-stroke pressure  
mid-stroke area covered  
direction of end-to-end line  
average direction  
start \$x\$  
stop \$x\$  
start \$y\$  
average velocity  
stroke duration  
stop \$y\$

#### **F1 scores**

**SVM with RBF: 0.878048780488**

**Log Reg: 0.780487804878**

When we added our two new features, the new observed performance is:

#### **F1 scores**

**SVM with RBF: 0.878048780488**

**Log Reg: 0.780487804878**

2.) The second features selection method tried to capture independent directions in the feature space. The first feature was selected using highest mutual information as the criterion. The rest of the features were sequentially chosen such that that each feature had minimum correlation with the features already chosen: The observed performance were:

mid-stroke pressure  
start \$x\$  
average direction  
20%-perc. dev. from end-to-end line  
inter-stroke time  
median acceleration at first 5 points  
20%-perc. pairwise velocity  
mid-stroke finger orientation  
20%-perc. pairwise acc  
mid-stroke area covered

#### **F1 scores**

**SVM with RBF: 0.878048780488**

**Log Reg: 0.756097560976**

When we added our two new features, the new observed performance is:

#### **F1 scores**

**SVM with RBF: 0.878048780488**

**Log Reg: 0.707317073171**

3.) The 3<sup>rd</sup> feature selection method trained a simple multi-class tree classifier to capture the variable importance for each feature. The 10 most important variables chosen were used for classification. The observed features and the corresponding F1 scores are:

mid-stroke pressure  
mid-stroke area covered  
start \$x\$  
stop \$x\$

direction of end-to-end line  
start \$y\$  
stop \$y\$  
average direction  
average velocity  
length of trajectory

**F1 scores**

**SVM with RBF: 0.829268292683**

**Log Reg: 0.780487804878**

When we added our two new features, the new observed performance is:

**F1 scores**

**SVM with RBF: 0.853658536585**

**Log Reg: 0.80487804878**

4.) The 4<sup>th</sup> feature trained a simple multi-class softmax model with high L-1 penalty (used to find sparse feature vectors) to find the 10 features with the highest coefficients. The chosen features and the corresponding performance are:

mid-stroke pressure  
mid-stroke area covered  
ratio end-to-end dist and length of trajectory  
stroke duration  
average velocity  
20\%-perc. dev. from end-to-end line  
length of trajectory  
median velocity at last 3 pts  
phone orientation  
direct end-to-end distance

**F1 scores**

**SVM with RBF: 0.80487804878**

**Log Reg: 0.780487804878**

When we added our two new features, the new observed performance is:

**F1 scores**

**SVM with RBF: 0.853658536585**

**Log Reg: 0.80487804878**

The performance of the classifier when trained on all the features was:

**F1 scores**

**SVM with RBF: 0.951219512195**

**Log Reg: 0.780487804878**

When we added are two new features, the new observed performance is:

**F1 scores**

**SVM with RBF: 0.951219512195**

**Log Reg: 0.80487804878**

**BEST FEATURE SET:**

Looking at the above results and qualitatively analyzing the signals each feature is trying to capture, the best 10 features in the original features that should result in the best possible performance should be:

mid-stroke pressure  
mid-stroke area covered  
start \$x\$  
average direction  
20\%-perc. dev. from end-to-end line  
20\%-perc. pairwise velocity  
20\%-perc. pairwise acc  
inter-stroke time  
stroke duration  
start \$y\$

These features should be able to capture all the useful signals and their combinations should be able to account for the variations observed in most of the other features. These features were chosen from the list of features selected using the maximum information criterion and the minimum correlation criterion. We see that the features chosen using these 2 methods produce the same results for the SVM-RBF classifier, so the features in these lists should be either complementary or overlapping. Since the signals captured by these 2 lists seems qualitatively different we assume that they are complementary and proceed as such choosing the features that occur in both lists first and then the ones that seem to capture some different aspect of the stroke and are present in one of the lists.

We also observe that most of the times addition of our new features doesn't result in much change classifier output. This could be attributed to the fact that our features are highly correlated with some of the features and the signal present in our features is already captured by the other. As a result, despite being very good predictors of the user ids themselves, they do not add much to the list of already known features.