

# Physiological Data Analysis Report

Akash Narayana, Gowri Srinivasa, Mohan Sadashivaiah, Shanmukh Katragadda  
{akash.narayana, mohan.sadashivaiah, shanmukh.katragadda}  
@daimlerchrysler.com, genetechie@yahoo.com  
DaimlerChrysler Research And Technology India, Bangalore.

## Abstract

*Gender classification was done based on the principal components of the raw sensor data. Annotation classification has two entries, one based on raw-sensor-signal classification and the other based on the principal components of the raw data. Support vector machines were used for classification.*

## Data preparation

The sensor data on visual inspection was found to contain outliers, and these were removed from the data. Sensors, especially Sensor 3 to 8 contained outliers which were eliminated, care was taken sometimes to delete the complete context data in a session which had lots of contiguous outliers. Many a time, the outliers corresponded to some 3 contiguous points, in these cases only these three points were removed. By this process 11,305 vectors were removed.

It was felt that the generalization ability of a classifier would be compromised if we had skewed distribution of training samples for the classes; hence care was taken to have balanced sets for training. Typically the class with the least samples was matched with the same number of vectors from the other class. This was necessitated due to the highly skewed ratio of the gender data and also for each of annotation data (3004 vs the rest and 5102 vs the rest).

## Gender classification

Gender classification of the raw signals with the characteristic 1, characteristic 2 & the 9 sensor signals was quite good. It was soon noticed that characteristic 2 has major correlation with the gender, and in fact was highly contributing to the classification. Hence in all later experiments both characteristic 1 & 2 were not used.

Principal component of the sensor data was analysed. The idea was to reduce the sensor data from 9 dimensions to lower dimensions (redundancies removed)

PCA for various dimensions from 5 to 8 were tried, and finally a dimension of 5 was chosen for the competition. Note: At the point of sending an entry we felt based on the classification results, a reduced dimension of 5 is suitable. But, at the point of writing this report, we aren't so sure.

Classification of gender was done at each timestamp. Majority voting was done based on the classifications in a session. Information from characteristic 1 & characteristic 2 were used at this point. We assumed that characteristic 1 & characteristic 2 are unique for a user.

## **Annotation classification**

It was felt that not enough representations from characteristic 1 & characteristic 2 are present for annotation learning, hence only the sensor signals were used for classification.

Training involved learning

1. Context 3004 vs. the rest
2. Context 5102 vs. the rest.

These were done by separate networks/systems.

In each of these cases, there exist 'either cases' whose classification is not of importance; hence those data samples were removed from the data sets. The 'rest of context' class in both the cases have a lot of samples compared to that of the context class. And, since we wanted the two classes to be learnt to be balanced in-terms of number of vectors, we reduced the number of samples in the 'rest of contexts' class by ensuring all contexts are indeed present but their numbers are reduced.

Here classification of the context is made for every data sample. Note, that since we had two systems to learn the contexts separately, there were a few instances when both the systems identified a vector as a positive sample of class 3004 & 5102 respectively. In such cases, the re-judgement was in favour of the context with the higher probability (in terms of number of vectors of that context), in this case, in favour of context 5102.

### **Entry 1:**

Here, the raw-sensor data was used for classification. The data was fed to a Support Vector Machine (SVM Torch) to learn the classification.

### **Entry 2:**

Principal components of dimension 7 of the sensor data was used to train the SVM classifier. Here, the dimension 7 was chosen arbitrarily, since we were running short of time.

Note: During training for context 5102, SVM Torch was taking a long time to converge, hence the data vectors were reduced to around 8000 vectors in each class and trained.

Another algorithm that was tried but wasn't so promising and hence not submitted was that of modeling a sensor. Here the idea is that for each context there is an underlying relationship amongst the sensors themselves. So, how 'bout trying to express/predict/estimate a sensor value based on the rest of the sensors. If the predicted value is far from the desired value then probably it belong to a different context. We could have separate systems/networks learning to predict each (or some) sensor for a context. The networks are trained for function approximation to learn the underlying function. And, when data from other context is given the error between the estimated value and the desired value is high (some threshold needed here). Well, this algorithm was checked with context 3004 but the results were bad (We didn't have time to check on context 5102) But, still we believe the idea is good, and perhaps it can be implemented in some other form and extensively tested.

## Conclusion

Using physiological signals to identify the activity and gender of a person is an interesting problem. We were interested in seeing if the minute-by-minute data clusters separately. Characteristic 1 (age) and characteristic 2 were not used for gender & annotation classification because it was felt that there were very few unique samples of them. Results of raw-sensor-data classified by SVM's were not so promising for gender classification (results not included in the report). Entries includes PCA based technique for gender classification and raw-sensor data and PCA based technique for annotation classification. The gender-classifier predicted the gender for each vector, the majority vote was used to assign the gender in a session. For annotation, separate systems learnt the contexts 3004 vs. the rest and 5012 vs. the rest, and gave their predictions for each minute-by-minute sensor values.