# Cluster Weighted Modelling: A modified mixture of Gaussians applied to analysis of biometric data

Matthew Gray

December 12, 1997

## 1 Problem Overview

From September 22nd to 25th, 1997, seven soldiers at the Fort Benning army base were equipped with sensor packages which measured a number of biometric and situational parameters: GPS (position, elevation, heading and speed), heartrate, core body temperature, and step count. They engaged in a number of activities, including a 20km road march, rest periods, and a number of assaults on training targets.

These data were sampled at a rate of 1 sample every 5 seconds, continuously, using the "Marathon Man" rig developed by the Personal Information Architecture group at the MIT Media Lab.

Given this biometric data, it is desirable to be able to discern the state of the soldier, especially to monitor for anomolies that might warrant medical attention, and to be able to automatically evaluate what activity a soldier is engaged in. Given the sample data, the second problem was deemed more immediately approachable.

Examining the data available by hand made it clear that a resting/working discriminator would be a trivial linear discriminant based on speed, as most of the work activity involved marching, and the rest activity involved little movement. The goal was to therefore get a attacking/not-attacking discriminator.

## 2 Model Overview

The model chosen is a "Cluster Weighted Modelling", a modified mixture of Gaussians system described by Neil Gershenfeld.

Each Gaussian($c_m$) attempts to predict the probability of a particular state, $s$ (attacking or not) and a particular vector $\vec{x}$ in the input space. This is factored as:

$$p(\vec{x}, s) = \sum_{m=1}^{M} p(s|\vec{x}, c_m)p(\vec{x}|c_m)p(c_m) \tag{1}$$

Seperable Gaussians are used. Each Gaussian may predict any state, however, for speed purposes, and to avoid certain pathological cases, some Gaussians was fixed to predict a particular state. That is, $p(s_m|\vec{x}, c_m) = 1$.

The Gaussians are initialized with random initial means in the range of the input space and variances equal to that of the input sample. Using E-M, the posteriors are calculated:

$$p(c_m|\vec{x}, s) = \frac{p(\vec{x}, s|c_m)p(c_m)}{\sum_{m=1}^{M} p(\vec{x}, s|c_m)p(c_m)} \tag{2}$$

The probability for each Gaussian to predict a given state (except the subset which are fixed to predict a particular state) is:

$$p(s|\vec{x}, c_m) = \frac{1}{Np(c_m)} \sum_{n|s_n=s} p(c_m|s_n, \vec{x}_n) \tag{3}$$

Similarly, the means, variances, and $p(c_m)$'s are updated. Given this, the forward values, $p(\vec{x}, s)$ are calculated and the process repeated.

## 3 Results

To confirm the system was working, a working/resting dataset was run through it, and, as expected, it yielded a discriminator based primarily on velocity. Though not visually apparent, heartrate data was also a substantial component of this discriminator.

The training data for the attack/non-attack classifier was composed of 3400 points encompassing one and a half days of activity for a single soldier, including extended marching periods, resting periods and an assault on a training target. Approximate times for the beginning and end of each of these activities had been logged and were used to tag the two states for the training data. This model was run with 15 Gaussians on a smoothed four-dimensional input space composed of horizontal speed, vertical velocity, heart rate, and body temperature.

An out-of-sample data set from a third day of excercises composed of 1900 points encompassing a full day including two training assaults and intermediate rest and marching periods was used for cross validation.

As seen in Figure 1, after the second iteration the out-of-sample error was minimized. As expected, errors on the in-sample data were extremely low, and the errors on the out-of-sample data were reasonable.

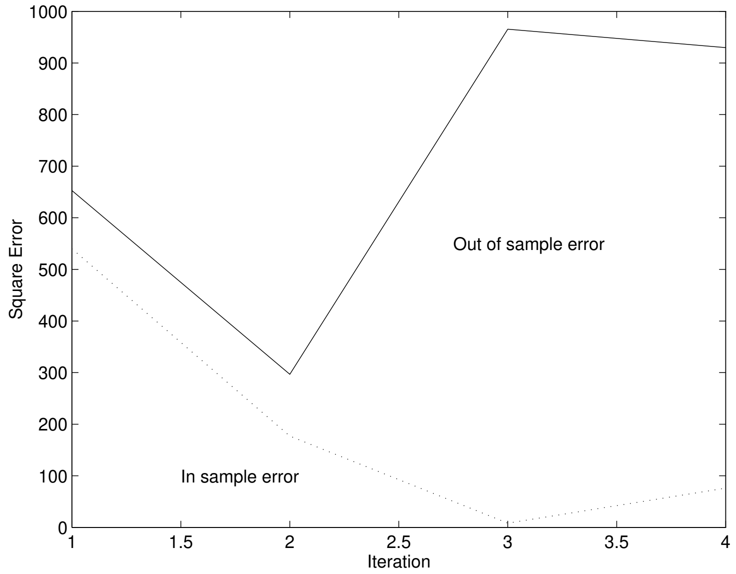


Figure 1: In and out of sample errors

Figure 2 shows a graph of the estimated probability that the soldier was engaged in an assault scenario (blue), along with the actual probability (red). Looking at this graph, it appears that the first assault extended roughly twelve minutes past the logged end time, and the second assault started roughly ten minutes before the logged start time. Both of these observations would be consistent with errors in the event log. The person who recorded the log indicated that the actual times may be as much as 20 minutes off from the recorded values.

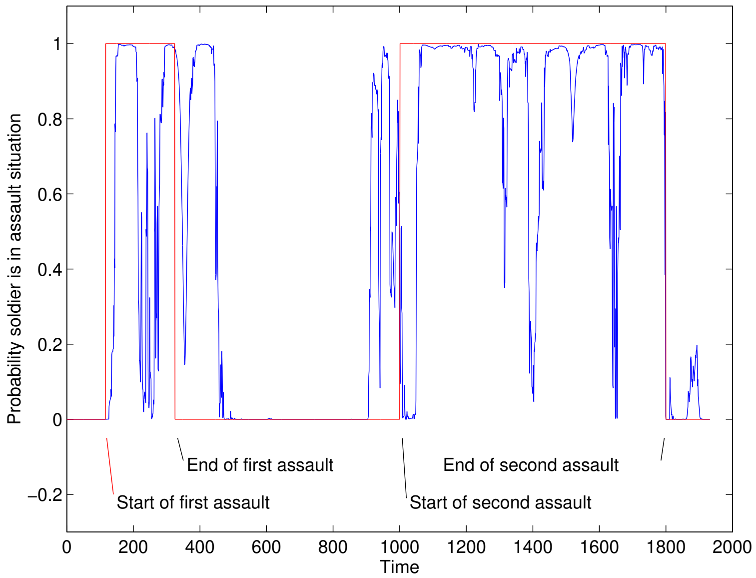


Figure 2: Probability soldier is currently engaged in an assault scenario

Overall, the discriminator performed very well. Further improvements could be made in performance, input space representation, more accurate collection of event logs, and using covariances instead of seperable Gaussians..