Sport sciences through data sciences

Keywords : Machine Learning; Sport.

Dimitri de Smet, Michel Verleysen, John A. Lee

Abstract – Machine learning aims at modeling relationships or patterns in data that can not be derived from domain-experts equations. This modelisation can be used to make predictions or to improve the domain-expert knowledge itself. The purpose of this project is to tackle sport science problems using machine learning techniques, taking advantage of the relatively new fact that zillions of geolocalized tracks are made publicly available on the web (Garmin, Nike+, Strava, RunKeeper, . . .)

Available tracks are tuples of geolocalized points associated with timestamps recorded by athletes with specific devices that can eventually record more parameters like heart rate, cycling torque, cadence, accelerations, temperature or barometric pressure.

We summarise hereafter four parts of this project.

Heart Rate Adaptation to Athlete's Power Output Many tracks are provided with continuous heart rate monitoring. Tracks allow estimation of the athlete's continuous power output. We can then identify the relationship $f()$ between Heart Rate $HR(t)$, and Power Output $PO(t)$.

$$HR(t+1) = f(PO(t)) \tag{1}$$

The expected benefits of this part are as follows:

- Improvement of the general model of heart rate adaptation to athletes' power output with a sample size exceeding current domain specific studies.
- Possibility of continuous monitoring of athletes' fitness level at minimal cost by parsing their activities

Difficulty of Routes A good intuition of the difficulty of a route (a tuple of geolocalized points without timestamps) is given by the average pace ($pace = k \times speed^{-1}$) atheletes can achieve on it. As routes were not run by the same set of athletes, we have to solve two coupled problems : evaluating the routes, and evaluating the athletes. We propose to do this with the help of a matrix completion algorithm (without going into details, the problems seem at first sight quite similar to the recommendation problem, which is well known in the machine learning literature). This will give us difficulty ratings for the routes that we can then use to fit a regression model $d()$ that takes the routes' features (distance, total ascent, ground type, ...) as inputs. The model can thereafter serve as objective measure of the difficulty of new routes that were not run yet .

$$\text{difficulty} = d(textroute) \tag{2}$$

The expected benefits of this part are as follows:

- Proper evaluation of sport routes (running, cycling, ...) helps athletes to prepare specific events and to optimize their pace on race days.
- An objective measure of the difficulty of a race would help organizers to establish new routes or to weight different races of a championship.
- It would allow comparing different runners performances achieved on different routes.

Workouts Assessment and Planning Access to all sport activities of given athletes gives us continuous monitoring of their fitness level. We can then build a more complex model that describes how the fitness level evolve with respect to workouts.

The expected benefits of this part are as follows:

- Helping workouts planning optimization to reach best fitness level at given events
- Automated adaptive workout planning

Data Preprocessing Gps traces contain noise, partly because of the devices technology (poor elevations resolution and need for continuous satellites visibility) and because of users misuses. These problems will first be addressed by prepossessing the data. For instance we will deal with elevation correction, smoothing, cropping and discarding tracks.

Figure 1: Louvain-La-Neuve Running Heatmap.

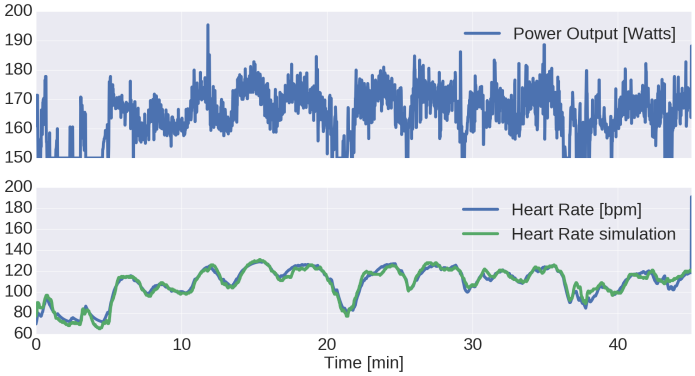

Figure 2: Heart Rate Simulation.