Running Race Times Prediction and Runner Performances Comparison using a Matrix Factorization Approach

Dimitri de Smet[1], Michel Verleysen[1] and Marc Francaux[2]
[1]*ICTEAM, Université catholique de Louvain, Louvain-la-Neuve, Belgium*
[2]*IoNS, Université catholique de Louvain, Louvain-la-Neuve, Belgium*

Keywords: Race Time Prediction, Athlete Evaluation, Athlete Comparison, Running, Collaborative Filtering, Matrix Factorization.

Abstract: This work provides tools based on matrix factorization that can be used to predict athlete running race times based on known race results. This is of interest for athlete preparation, for workout route planning and for race events organization. This work differentiates from previous ones by jointly considering athletes and routes. This work shows how race records can be used to infer knowledge on the users and the races. The same tools can also serve to compare different athlete performances and track athlete level over time. Experiments were conducted on race records of 648 athletes from casual to elite levels. Experiments show that the methodology can be applied to real data and gives relevant insights.

1 INTRODUCTION

More and more athletes record their sport activities using a smartphone or a sport watch. Their records are stored on remote servers owned by companies that, usually, provide health or fitness statistics. The massive amount of tracks that are recorded every day forms a basis for learning about sport practice, locations and athletes.

This paper shows how to take benefit from geo-localized running tracks to build models that can be used to predict running race times. For this purpose, this paper presents a methodology that allows the characterization of any route in order to be able to predict running race times of an athlete for which a few race records are known. This is of particular interest for race preparation and for workout route planning. To our knowledge, previous research on race time prediction focus on the athletes' features (age, morphology, gender, ...), the athletes' training measurements (Noakes et al., 1990; Vickers and Vertosick, 2016; Rüst et al., 2011) or the routes' characteristics (Riegel, 1981). The framework that is proposed here allows to capture complex relationships that account for athletes and routes at the same time. Moreover, the methodology provides tools that enable athlete performances comparison that can be used to compare different athletes or to track individual progression over time.

In this work, only tracks recorded during race events are considered because they are more likely to reflect athletes' abilities than casual activities taken at random. The task of comparing races based on race records is not trivial because races are not attended by the same set of athletes. Likewise, athletes' results cannot be compared directly because athletes do not always attend the same races. On the other hand, there exist some points of comparison because certain races have some athletes in common. Making the most of these points of comparison is related to a research area that is called 'collaborative filtering' which includes the well-known problem of product recommendation (Ekstrand et al., 2011).

Experiments were performed on a small set of races that took place in a restricted geographic area to favor points of comparison (athletes who share several races). Our database contains 648 users of a sport activities sharing platform. Users recorded 276 races over a period of two years.

2 METHOD

The methodology is divided in two main processes that are validated individually.

First, a collection of race results is used to infer a few underlying variables for each athlete and for each

de Smet D., Verleysen M. and Francaux M.
Running Race Times Prediction and Runner Performances Comparison using a Matrix Factorization Approach.
DOI: 10.5220/0006499300960101
In *Proceedings of the 5th International Congress on Sport Sciences Research and Technology Support (icSPORTS 2017)*, pages 96-101
ISBN: 978-989-758-269-1

race using matrix factorization (as it will be explained in section 2.1). These athlete and race variables will be referred to as athlete vectors and race vectors. It is assumed that they synthesize the information that is required to predict race times. Athlete vectors $\mathbf{a}_a$ and race vectors $\mathbf{r}_r$ are purely abstract and have no immediate physical counterpart.

Unlike athlete vectors, race vectors are related to objective physical quantities that are, for instance, linked to the topology of the race. The second process establishes a relationship between the elements of race vectors and the actual race route characteristics $\mathbf{x}_r$, like total distance or cumulative elevation gain. Therefore, a regression model $F(.)$ is built such that $F(\mathbf{x}_r)$ models as accurately as possible the $\mathbf{r}_r$ vector produced by the first process.

The combination of the two processes allows race time prediction on routes that were not yet run and for which only the itinerary is known. Indeed, route characteristics $\mathbf{x}_r$ can be computed from the itinerary; race vectors can be obtained using the regression model $\mathbf{r}_r \approx F(\mathbf{x}_r)$ and race time for athlete a running race r can be predicted from athlete vector $\mathbf{a}_a$ and race vector $\mathbf{r}_r$.

2.1 Matrix Factorization

A matrix of race times $(t_{a,r})$ with columns corresponding to races and rows corresponding to athletes is filled with known race times. In practice, most matrix entries are unknown (over 90%) because most athletes ran only few of the existing races. The matrix factorization will allow to make predictions for the missing values.

Race times $t_{a,r}$ are normalized with respect to race distances d_r to give the race average paces

$$p_{a,r} = \frac{t_{a,r}}{d_r}$$

which have better statistical distribution properties and intuitively correspond to the notion of race difficulty: the faster they are run, the easier they are. The normalization also gives the advantage that the error criterion detailed below will not favor long races.

The matrix factorization depicted in Figure 1 results in an abstract vector representation of chosen length N for each race r and for each athlete a, respectively $\mathbf{r}_r$ and $\mathbf{a}_a$. The vectors summarize and smooth the information contained in the known race results. It is postulated that the vectors can then serve to generate predictions for the missing race results.

2.1.1 Model

As shown in Figure 1, it is assumed that the average race paces depend linearly on both race vectors $\mathbf{r}_r$ and

$$\overbrace{\text{athletes}\left\{\begin{bmatrix} p_{0,0} & \cdot & \cdot & p_{0,3} & p_{0,4} & p_{0,5} \\ \cdot & p_{1,1} & \cdot & p_{1,3} & \cdot & p_{1,5} \\ \cdot & \cdot & p_{2,2} & \cdot & p_{2,4} & p_{2,5} \\ p_{3,0} & \cdot & \cdot & p_{3,2} & p_{3,3} & \cdot \\ p_{4,0} & p_{4,1} & p_{4,2} & \cdot & \cdot & \cdot \\ \cdot & \cdot & p_{5,2} & p_{5,3} & \cdot & p_{5,5} \\ \cdot & p_{6,1} & p_{6,2} & \cdot & \cdot & p_{6,5} \\ p_{7,0} & p_{7,1} & p_{7,2} & \cdot & \cdot & \cdot \end{bmatrix}\right.}^{\substack{\mathbf{P} \\ \text{races}}} \approx$$

$$\underset{\substack{\mathbf{A} \\ \text{(athlete vectors)}}}{\begin{bmatrix} a_{0,0} & a_{0,1} & a_{0,2} \\ a_{1,0} & a_{1,1} & a_{1,2} \\ a_{2,0} & a_{2,1} & a_{2,2} \\ a_{3,0} & a_{3,1} & a_{3,2} \\ a_{4,0} & a_{4,1} & a_{4,2} \\ a_{5,0} & a_{5,1} & a_{5,2} \\ a_{6,0} & a_{6,1} & a_{6,2} \\ a_{7,0} & a_{7,1} & a_{7,2} \end{bmatrix}} \times \underset{\substack{\mathbf{R}^\intercal \\ \text{(race vectors)}}}{\begin{bmatrix} r_{0,0} & r_{1,0} & r_{2,0} & r_{3,0} & r_{4,0} & r_{5,0} \\ r_{0,1} & r_{1,1} & r_{2,1} & r_{3,1} & r_{4,1} & r_{5,1} \\ r_{0,2} & r_{1,2} & r_{2,2} & r_{3,2} & r_{4,2} & r_{5,2} \end{bmatrix}} = \tilde{\mathbf{P}}$$

Figure 1: The matrix factorization produces a vector of length N (here $N = 3$) for each race and for each athlete. Dots represent unknown entries for which predictions can be achieved by dot product.

athlete vectors $\mathbf{a}_a$. Average paces $p_{a,r}$ can therefore be expressed as a sum of N product terms

$$p_{a,r} = \sum_{i=0}^{N-1} a_{a,i} \cdot r_{r,i} + \varepsilon_{a,r} \tag{1}$$

where ε refers to an added noise that reflects the fact that our relationship is not deterministic and only holds true on average. Predictions can be expressed by

$$\tilde{p}_{a,r} = \sum_{i=0}^{N-1} a_{a,i} \cdot r_{r,i}. \tag{2}$$

2.1.2 Optimization

The problem to be solved is to choose athletes and races vectors that best reproduce observed average race paces. Let Ω be the set of observed entries (a,r) of the matrix $p_{a,r}$. Using the least square error criterion, athlete and race vectors can be found by solving the optimization problem

$$min_{\mathbf{A},\mathbf{R}} \sum_{(a,r)\in\Omega} (p_{a,r} - \mathbf{a}_a \cdot \mathbf{r}_r^\intercal)^2. \tag{3}$$

This is a non-convex problem that can be solved using heuristics that proved to converge well in practice. Successful experiments are conducted using *stochastic gradient descent* and *alternate least square* algorithms. *Stochastic gradient descent* loops over Ω slightly modifying $\mathbf{A}$ and $\mathbf{R}$ in the direction of the steepest gradient; reducing the error criterion at each step (Gemulla et al., 2011). *Alternate least square* alternates between two convex problem: optimization of the **R** vectors while holding **A** vectors constant and optimization of the **A** vectors while holding the **R** vectors constant (Jain et al., 2013) .

2.1.3 Matrix Requirements and Choice of N

Obviously, the matrix factorization cannot produce sound vectors if the number of elements to estimate is larger than the number of known entries. This condition is met only if each race has at least N athlete results to estimate their N vector elements and if each athlete has at least N race results. Races and athletes that do not fulfill this condition are removed from the database. In practice, as race times account for variables that do not come into picture here (such as athlete's fitness of the day or weather conditions), N vector elements require slightly more than N race results so that the noise on the data can be averaged out.

Parameter N is related to the complexity of the model. A small N value corresponds to a simple model that is more likely to generalize well but that might not capture the entire information included in the race results matrix. A larger value of N requires more data.

Choosing $N = 1$ would mean that the model assumes that a race time only depends on one parameter per race (that can be interpreted its difficulty) and one parameter per athlete (that can be interpreted as the inverse of his fitness level).

Choosing $N > 1$ allows a multivariate representation of what makes races faster or slower and a multivariate representation of corresponding athlete abilities. For instance, if vector elements can be mapped to route characteristics, the model could express the race difficulty and athlete abilities in terms of endurance, ascent or ground surface type.

2.1.4 Matrix Factorization Validation

To quantify the ability to determine the **A** and **R** vectors that approximate the known race results, 10 percent of the known matrix entries are removed from the initial set (prior to matrix factorization) and then compared to the same entries in the approximated paces matrix $\tilde{\mathbf{P}} = \mathbf{A} \cdot \mathbf{R}^{\intercal}$. The process is repeated 10 times, keeping 10 other percent apart. The accuracy is then taken as the average root mean square error over the 10 repetitions. This process corresponds to a 10-fold cross validation scheme (James et al., 2013).

2.2 Mapping Race Vector Elements to Objective Route Features

The aforementioned matrix factorization gives vectors of N athlete variables and vectors of N race variables. As it would be of practical use to estimate route vectors on new routes for which race results are not (yet) known, the known route parameters $\mathbf{r}_r$ need to be linked with objective route features.

2.2.1 Route Features

Routes are described as a list of geo-localized coordinates to which corresponding altitudes can be associated. Although one can argue that other parameters can be related to race times (such as weather conditions and ground type), our characteristics are solely based on the elevation profile (such as the one in Figure 2) because they can be easily gathered based on Global Positioning System (GPS) positions present in track files. As consumer grade GPS-based elevations cannot be trusted (Bauer, 2013), elevations were obtained using *google map APIs* queries and then smoothed with a 200 meters moving average filter. This process was validated with barometric altimeters.

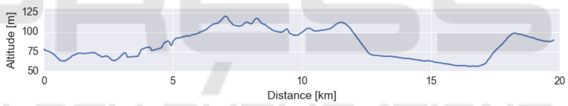

Figure 2: Elevation profile example that serves as input to generate route features.

Experiments were only conducted with the two most common route features, namely *total distance* and *cumulative elevation gain*. The later represents the sum of positive vertical distances taken along the track.

2.2.2 Regression Model

Let **X** be the route features matrix in which each row $\mathbf{x}_r$ corresponds to the route features of race r. A regression model $F(.)$ is built to map the race vector elements to the race features:

$$\tilde{\mathbf{r}}_r = F(\mathbf{x}_r). \tag{4}$$

In the present work, the function $F(.)$ is chosen to be a multiple output linear regression.

2.2.3 Regression Validation

To validate the regression model, a fraction of the races are removed from the initial matrix. Their vectors are estimated with the introduced regression model using (4). Race paces are predicted with (2) dot products. The matrix factorization and the regression model are jointly validated by comparing these prediction with actual race records that were kept aside, exactly as it was performed for the matrix factorization validation.

3 USE CASES

Two cases are described here to illustrate how the methodology exposed in this paper can be used in practice. In the first case, the aim is to predict the running time of an athlete for a race that he did not run (yet). The second case is the problem of athletes' performance comparison based on running times that were recorded on different routes.

3.1 Race Time Prediction

In the simplest case, where the athlete and the race vectors are known from the matrix factorization, race pace can be predicted using (2). Otherwise, vectors can be estimated as follows:

- If the race is new, the race vector $\mathbf{r}_r$ is not known but it can be computed using (4) starting from its actual route properties.
- If the athlete was not in the initial race results matrix, his vector $\mathbf{a}_a$ can be computed using a set of, at least, N known race paces $\mathbf{p}_{\Omega_a}$ by solving a multiple output linear regression for $\mathbf{a}_a$:

$$\mathbf{R}_{\Omega_a} \cdot \mathbf{a}_a^\mathsf{T} \approx \mathbf{p}_{\Omega_a} \tag{5}$$

for which the race vectors $[r_{r,1}, r_{r,2}, \ldots, r_{r,N}]$ in $\mathbf{R}_{\Omega_a}$ are again either known from the initial matrix factorization or computed using (4).

3.2 Athletes Comparison

Athlete vectors are either known from the matrix factorization or computed by solving (5). If the athlete vector elements thus obtained can be mapped to some actual athletes' abilities, athlete's vectors can be used directly for athlete comparison. Otherwise, the race prediction method can be used to simulate performances on chosen races. Athletes can therefore be compared through their simulated race times on benchmark race routes.

Table 1: Initial dataset characteristics based on the choice of minimum number of races per athlete : number of races, of athletes, of race times and proportion of observed race times with respect to the matrix size.

Min. race per athlete	Races	Athletes	Races Times	Observed [%]
1	276	648	2990	1.7
10	251	67	1010	6.0
15	199	25	521	10.5
20	145	12	308	17.7
25	76	6	182	39.9
30	65	3	99	50.8

4 RESULTS

Results obtained with an initial database containing 2990 race times of 648 athletes on 276 races are discussed separately for the matrix factorization (section 4.1) and then for race time prediction on new races using the regression model (section 4.2).

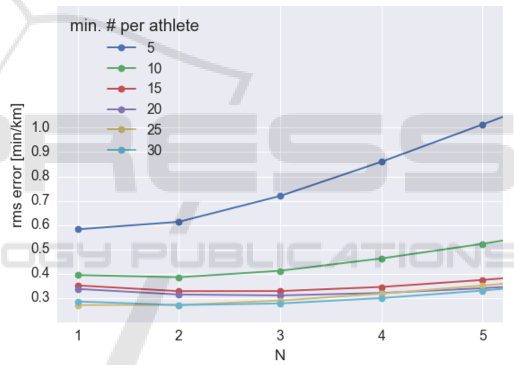

Figure 3: The matrix reconstruction root mean square error depends on the number of races per athlete and the model parameter N.

4.1 Matrix Factorization Results

The ability to predict race times based on matrix factorization depends on the available data and on the algorithm parameter N (see Figure 3). As shown in Table 1, the initial database can be pruned based on the requirement on the minimum number of races per athlete. The higher this value is, the more athletes are removed from the database and the easier it is to predict race times.

The root mean square error on race pace prediction can be as low as 20 seconds per kilometer on average. It is significantly improved when only very active athletes are kept (over 10 races per athlete). Optimal model parameter N is found to be between 1 and 3 depending on the dataset. The uncertainty that might seem high accounts for the fact that some race times can highly differ from what was expected because of an injury or simply because the athlete did not try to achieve is best potential performance.

4.2 Results for Predictions on New Races

Race time prediction accuracy is measured for new athletes on new races. For this purpose, it is necessary to know at least N race times for each athlete and that the routes itinerary is known to compute their features (distance and cumulative elevation gain).

A few race results can be used to generate athletes vectors using (5) and route features can give race vectors through the regression model given by (4). Pace can then be predicted on known races, and compared to actual race records. In this case, race vectors are unknown and are estimated from route properties. Therefore, the root mean square error increases to about 26 seconds per kilometer. Figure 4 shows race pace predictions versus actual records.

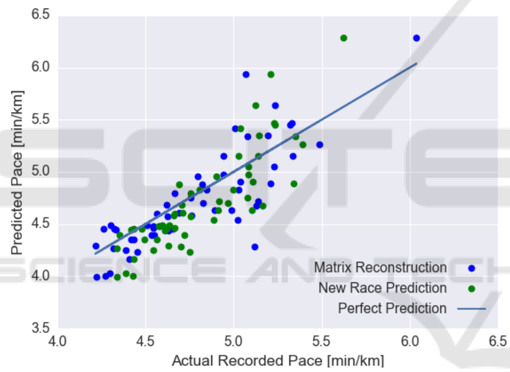

Figure 4: Predictions versus observed race times scatter plot with known vectors or with estimated ones.

5 CONCLUSION

This paper provides tools that can be used to predict race times. This is of interest for athlete preparation, for workout route planning and for race events organization. The same tools can also serve to compare different athlete performances and to track athlete level over time.

Experiments show that the methodology is applicable to real data and gives meaningful results. This work will be continued in different directions. First, only the two most commonly used route features were used (distance and cumulative positive elevation gain) but any function of the elevation profile could lead to better predictive performances. Other route parameters such as ground type and weather conditions may also prove to improve time prediction.

Then, race vector elements were assumed to be a linear function of the race features. Other nonlinear regression models could improve the accuracy as well. Two different approaches can be pursued.

First, domain knowledge was not considered in this work. Most probably, accuracy could benefit from well-established physiological or empirical models; for instance, the relationship between average race pace and distance has been modeled in other works by hyperbolic law, power law or nomogram (Péronnet and Thibault, 1989; García-Manso et al., 2012; Coquart et al., 2015).

A second path to be taken would be to discover more complex relationships between route features and race pace through the data itself using model fitting techniques. This approach will probably require a larger amount of race data.

REFERENCES

Bauer, C. (2013). On the (in-) accuracy of gps measures of smartphones: a study of running tracking applications. In *Proceedings of International Conference on Advances in Mobile Computing & Multimedia*, page 335. ACM.

Coquart, J. B., Mercier, D., Tabben, M., and Bosquet, L. (2015). Influence of sex and specialty on the prediction of middle-distance running performances using the mercier et al.s nomogram. *Journal of sports sciences*, 33(11):1124–1131.

Ekstrand, M. D., Riedl, J., and Konstan, J. A. (2011). Collaborative filtering recommender systems. *Foundations and Trends in Human-Computer Interaction*, 4:175–243.

García-Manso, J., Martín-González, J., Vaamonde, D., and Da Silva-Grigoletto, M. (2012). The limitations of scaling laws in the prediction of performance in endurance events. *Journal of theoretical biology*, 300:324–329.

Gemulla, R., Nijkamp, E., Haas, P. J., and Sismanis, Y. (2011). Large-scale matrix factorization with distributed stochastic gradient descent. In *KDD*.

Jain, P., Netrapalli, P., and Sanghavi, S. (2013). Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning. volume 112, chapter 5, pages 176–186. Springer.

Noakes, T. D., Myburgh, K. H., and Schall, R. (1990). Peak treadmill running velocity during the v o2 max test predicts running performance. *Journal of sports sciences*, 8(1):35–45.

Péronnet, F. and Thibault, G. (1989). Mathematical analysis of running performance and world running records. *Journal of Applied Physiology*, 67(1):453–465.

Riegel, P. S. (1981). Athletic records and human endurance: A time-vs.-distance equation describing world-record performances may be used to compare the relative endurance capabilities of various groups of people. *American Scientist*, 69(3):285–290.

Rüst, C. A., Knechtle, B., Knechtle, P., Barandun, U., Lepers, R., and Rosemann, T. (2011). Predictor variables for a half marathon race time in recreational male runners. *Open access journal of sports medicine*, 2:113.

Vickers, A. J. and Vertosick, E. A. (2016). An empirical study of race times in recreational endurance runners. *BMC Sports Science, Medicine and Rehabilitation*, 8(1):26.