Pacing Strategies and Athletic Performance in Long-Distance Running

Arie-Willem de Leeuw[1], Laurentius A. Meerhoff[1], and Arno Knobbe[1]

Leiden Institute of Advanced Computer Science (LIACS), Niels Bohrweg 1, Leiden, the Netherlands
a.de.leeuw@leidenuniv.liacs.nl

In this work, we have investigated the effect of age, gender, and pacing properties on the performance in long-distance running. We have used public data of races on the 10 km, the half marathon, and the full marathon organized by the Boston Athletic Association. The number of participants varied per distance and gender, with 9,464 male and 12,189 female on the 10 km; 8,480 male and 10,205 female on the half marathon; 43,125 male and 35,906 female on the full marathon. In addition to the final times, we have two intermediate times in the 10 km and half marathon races. The data of the marathon contains intermediate times after every 5 km and halfway the marathon.

First, we have developed distance and gender-specific models for describing the relationship between age and performance. To obtain the model that is the best description of this complex dependence, we consider polynomials of arbitrary degree m. For a choice of m, we apply stratified sampling by sorting the data based on the age of participants and randomly distribute every 10 successive datapoints over 10 different sets. Hereafter, we apply 10-fold cross-validation, perform least squares regression, and select the degree m that has the minimal value for the sum of squares of errors on the validation set. Finally, the value of the corresponding coefficients of the polynomial of degree m is then obtained by using ordinary least squares regression on the complete dataset.

Secondly, we focused on pacing strategies as this is considered one of the most important factors for a successful long-distance run [1]. To compare athletes that run at different average speeds, we defined the relative pace as the average speed between two successive intermediate points divided by the average speed during the entire race. On every distance, we have collected the relative paces of the runners to obtain athlete-specific speed profiles. Hereafter, we have applied k-means clustering to all profiles to identify the three most characteristic pacing strategies on all distances. We also have found relationships between the final time and the followed pacing strategy.

Finally, we have used an exploratory data mining paradigm, i.e., Subgroup Discovery [2], to find the characteristics for having the best possible performance in a race. We have constructed a dataset in a tabular format, where the columns represent the variables, or in Subgroup Discovery-terminology *attributes*, that characterize the pacing throughout the race, such as the speed change from the first to the second half of the race. Since Subgroup Discovery is a supervised technique, each row also contains information about the target variable, which in our case is the relative difference between the final time of an athlete and the time that is predicted by the age-performance model.

We have used the Cortana Subgroup Discovery tool [3] to select the pacing properties that have the largest impact on the performance. We have found that controlling the pace changes is the most important feature for optimizing the performance. For example, for the marathon, we have found that men on average perform 7.30% better than the model predicts, if the interquartile range of the pace changes is smaller than 7.66% and the deceleration from 0 – 20 km to 20 – 30 km is maximally 11.5%. For women, we have obtained that runners on average perform 6.14% better than the prediction of the age-performance model, if the interquartile range of the pace changes is smaller than 8.07% and the maximum relative pace is smaller than 1.13. Thus, for both men and women, the middle 50% of the pace changes should fall within roughly 8% of each other. Moreover, this shows that pacing has a large impact on the result in long-distance running events, and thus besides optimizing the fitness of an athlete, is probably one of the biggest factors in running performances.

The results we obtained in this research give concrete and relatively simple conditions about the right way to approach a running challenge. However, since the expected improvements are based on the population, these trends may not hold for the individual [4]. Therefore, for future research, it would be interesting to collect data of multiple races of individual athletes. With the methods that are used in this research, we could give an athlete personalized advise about his or her ideal pacing strategy and investigate to what extent results depend on personal characteristics, such as weight or fat percentage.

Keywords: Sport Analytics; Running; Age-Related Performance; Pacing Strategy; Regression; Subgroup Discovery

References

1. Abbiss C.R, Laursen P.B. Describing and understanding pacing strategies during athletic competition. Sports Medicine 2008; 38(3):239 – 52
2. Novak P.K, Lavrač N, Webb G.I. Supervised descriptive rule discovery: a unifying survey of contrast set, emerging pattern and subgroup mining. J Mach Learn Res 2009; 10:377 – 403
3. Meeng M, Knobbe A. Flexible enrichment with Cortana – software demo. BeneLearn, the annual Belgian-Dutch conference on machine learning 2011: 117 - 119
4. Barr D.J, Levy R, Scheepers C, Tily H.J. Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language 2013; 68(3): 255 – 278.