Using Machine Learning Techniques to Support Marathon Runners

Ciara Feely

Science Foundation Ireland Centre for Research Training in Machine Learning, University College Dublin

Abstract

This document describes the research plan for the project "Using Recommender Systems Techniques to Support Endurance Athletes, in particular Marathon Runners". My PhD began in September 2019 and I am due to complete it in 2023, which means the next 12 months will be focused on completing the remaining research and preparing a thesis for submission. Building on work previously presented at ICCBR, this project aims to develop supports for runners including performance prediction, understanding of the risk of sustaining training disruptions, and recommending tailored training plans by applying machine learning techniques to the noisy, inconsistent time series data that is routinely collected by wearable sensors. The progress consists of extracting weekly training features from 15 million sessions for 300,000 unique marathon runners. Several publications involving race-time prediction, training plan recommendation, and training disruptions have been produced, so far working on a reduced dataset. Future work planned will include further feature engineering, applying the models to the larger dataset and providing a more detailed evaluation of training recommendations and explanations with a user study.

Keywords

CBR for health and exercise, marathon running

1. Research Problems

The purpose of this PhD is to develop a recommender system that generates tailored training programmes that enable marathon runners to achieve their performance goals using machine learning with the raw activity data that is routinely collected by wearable sensors. A recommender system has several components. First we must elicit a user's preferences to know how a given recommendation will benefit them. Here that translates to converting the raw activity data that comes in the form of time-series tracked at regular intervals for distance, speed, and elevation, into a suitable level of abstraction – daily, weekly, and monthly features, allowing for several machine learning models to be compared and contrasted to predict marathon performance based on training. Second we need to understand the cost of the recommendations, and for marathon running this would be how under or overtraining or training inconsistently could cause a runner to sustain an injury and miss their race or simply not meet their full potential. Finally we need a way to represent suitable options to the user i.e. to present runners with various training programmes and explaining the benefit and potential cost that following or

ciara.feely@ucdconnect.ie (C. Feely)

CEUR Workshop Proceedings (CEUR-WS.org)

not following the programmes has on their performance. The research questions cover each of these.

1. Can we predict marathon performance from activity data?
2. Can we model training consistency and the risk of sustaining training disruptions?
3. Can machine learning be used to recommend tailored training programmes to marathon runners?

1.1. Feature Engineering and Performance Prediction

To plan their training, marathon runners need an accurate estimate of their fitness and future marathon performance. The sports science community has generated dozens of different techniques for estimating fitness and predicting performance on race-day based on key physiological fitness indicators, training programme summaries and race histories, however there is no single approach that works well for runners of all backgrounds [1]. Previously, in Smyth and Cunningham [2, 3] the authors used a CBR model to predict future personal best time based on previous race-times with recreational runners. While this was accurate, it was not a suitable method for novice runners who do not have previous marathon experience. What is lacking is a marathon performance prediction model that can produce predictions for elite and recreational runners alike, adaptable to each point in training. GPS and heartrate data routinely captured by wearable sensors allows for more detailed features such as fastest 10km and average heartrate to be calculate per individual training sessions, and this PhD will involve the creation of a variety of representations of such data to facilitate machine learning tasks, for example extracting daily, weekly or monthly features. Many of these features will not be explicit in the data, for example features that capture a runner's recovery time, and it is also possible that different representations and models will be required for different runners for example males versus females, or recreational versus elite runners.

1.2. Understanding Training Consistency and Disruption Risk

When developing a recommender system that generates training programmes to improve a marathoner's performance, the potential cost for a runner adopting the recommendation is that the programme might be too difficult, causing them to sustain some sort of injury, or too easy causing them to train and perform sub-optimally. Running is an extremely taxing sport, with an estimated 2.5 times a runner's bodyweight hitting the ground with each stride. Novice runners and longer distance runners such as marathoners are at the greatest risk of sustaining an injury [4]. While a number of different causes of injury have been studied – from training load features such as total weekly distance, to physiological variables related to running gait, to personality traits – there has been no concrete cause found other than that having a history of running related injuries makes you more likely to have a subsequent injury. *Consistency* in training is a key marker that the runner's training plan is at the appropriate level – striking the balance between intense and achievable such that they will be able to optimise their performance. A *training disruption* on the other hand is a marker that the training could be too intense potentially leading to some running related injury, or that the runner has become demotivated. Regardless of the reason for their inconsistency in training runners will need to understand how this has impacted their performance and how they can get their training back on track following some disruption.

1.3. Recommending and Explaining Training Plans

Training for a marathon requires at least 12-16 weeks of intense training with sessions of a variety of goals to build strength, endurance, and speed. While elite marathon runners might have access to coaches to help them plan their training programmes, recreational runners rely on one-size-fits-all training programmes gleaned from an internet search. These programmes use goal marathon pace, but otherwise these programmes are not tailored to the runner's current status, history, training preferences, lifestyle etc. Additionally as training progresses the training programmes do not adapt to become perhaps more ambitious if a runner is over-performing, or to become lighter if a runner has experienced some training set-back. This research question surrounds developing a recommender system capable of offering tailored training programmes to runners, that are adaptable as training progresses. Explaining the training suggestions and presenting them to runners suitably is imperative for trust and adoption of the system.

2. Research Plan

This research project commenced in October 2019, and the intended finish date is August 2023. In what follows I will give an overview of the work to date and proposed future work for each of the main research questions.

2.1. Feature Engineering and Performance Prediction

For my PhD I have access to an anonymised set of all sessions uploaded into popular mobile fitness application Strava between 2014-2017, via a data-sharing agreement. I have since identified over 500,000 unique runners who have completed marathon distances, and extracted training features for the 16 week programme leading up to race-day. Data cleaning and censoring has resulted in a set of 15 million training activities for over 500,000 marathon programmes from just under 300,000 unique runners.

Work to date consisted of building case-based reasoning models to predict marathon performance at any point in training using training features [5, 6], and also incorporating previous race-times [7]. The current best performing model achieves error rates of approximately 5-8% depending on the point in training. The planned work is to further explore features for performance prediction in particular to better summarise training sessions for example the variability of heartrate or pace might indicate an interval session. The incorporation of heartrate data could also give an indication of the level of intensity of individual training sessions. Another key step is to investigate methods beyond case-based reasoning for performance prediction – so far some baseline linear models and decision tree models have been tested but were found to be less accurate than the case-based reasoning models. An exploration of more black-box models is planned to understand the performance-explainability trade-off in this domain.

2.2. Understanding Training Consistency and Disruptions

One difficulty in working with raw activity data is that we lack information from runners – so we don't know what a runner's goal is, whether a runner is finding the training too challenging or too easy, whether they have become injured or what their race or injury history is. To combat this, to date the work on understanding the cost of running recommendations has focused on training disruptions – lengthy periods without any training activities – as a proxy for injury. We presented a model predicting training disruptions with 20,000 cases that lacked accuracy – approximately 60% accurate, however the use of a case-based reasoning model and counterfactuals allowed for a fully explainable output to runners to facilitate their understanding of why their risk of sustaining a disruption was high based on training patterns of other runners [8]. Presently I am conducting a large scale data analysis of training disruptions – their frequency, impact on marathon performance and how training leading up to and returning from training presents – which will be presented in a sports analytics journal. Future work will involve reapplying the previous prediction model with the larger dataset and improved feature engineering to hopefully improve the error rates, and additionally to combine this model with the performance-based training recommendations to help runners understand how pushing for a more ambitious goal may have some risks to consider.

2.3. Recommending and Explaining Training Plans

Thus far training recommendations have been incorporated into the marathon performance prediction model [5, 6] by filtering the case-base by goal-time and using the future weeks of training of the most similar runner to the query runner based on training to this current point. I used a similar approach to generate training recommendations based on reducing injury risk [8] and here counterfactuals were incorporated to help runners understand what specifically in their current training has contributed to them having a greater risk of experiencing a training disruption.

The planned future work is to generate more detailed training explanations based on performance goals, and to evaluate these further. Counterfactual explanations allow runners to reflect on things they could have done differently along the training programme, and prefactual explanations will offer runners a sense of what adaptions they can make to their training programme now to reach their goal. Up until now evaluation has been an offline proof-of-concept style evaluation, and a user-study is planned with runners to ascertain whether the training explanations and recommendations are sensible to runners. Additionally, I need to consider how this can be presented to runners. While the case-based recommendation system allows us to easily retrieve the suitable training features, however, presenting these features will not necessarily translate into runners knowing what training to complete. Thus a user study with mock-ups of training plans is planned to understand how to present this to runners.

2.4. Timeline

The aim is to spend the remainder of 2022 finishing up the research projects outlined above, and to then move onto writing the thesis ahead of submission in August 2023. An initial literature review has been written up, as well as a first draft of chapters describing the dataset and work to date.

2.5. Expected Contributions

The main contribution of this work is to develop a system capable of recommending an adaptable programme of training activities to help recreational marathon runners achieve their performance goals safely and effectively. The complexity comes from recommending a series of activities, compared to a one-shot recommended item, as well as from representing the noisy, inconsistent, and unlabelled sensor data available. I believe that the work will have benefits for the running community, and running related research, and that the representations and models developed in this PhD will be applicable to other endurance sports such as races of other distances, cycling, and swimming, and any other recommender systems domain that involve recommending an ordered series of items.

2.6. Conclusions

This PhD aims to develop supports for marathon runners as they train. A dataset of over 500,000 marathon training programmes has been extracted, and modelling for race-time prediction and injury risk based on training programmes completed and presented at conference proceedings. Future work involves finalising the research into feature engineering, and providing counterfactual explanations of training recommendations, as well as writing up the thesis. Being awarded the opportunity to participate at the ICCBR DC would enable my engagement with senior members and also students of the CBR community and receive useful guidance as I finalise my research plans in advance of my final year.

2.7. Acknowledgements

This work is supported by Science Foundation Ireland Centre for Research Training in Machine Learning (18/CRT/6183).

References

[1] A. Keogh, B. Smyth, B. Caulfield, A. Lawlor, J. Berndsen, C. Doherty, Prediction equations for marathon performance: A systematic review, International Journal of Sports Physiology and Performance 14 (2019) 1159–1169.

[2] B. Smyth, P. Cunningham, Running with cases: A CBR approach to running your best marathon, in: Case-Based Reasoning Research and Development - 25th International Conference, ICCBR 2017, Trondheim, Norway, June 26-28, 2017, Proceedings, 2017, pp. 360–374. doi:10.1007/978-3-319-61030-6_25.

[3] B. Smyth, P. Cunningham, An analysis of case representations for marathon race prediction and planning, in: Case-Based Reasoning Research and Development - 26th International Conference, ICCBR 2018, Stockholm, Sweden, July 9-12, 2018, Proceedings, 2018, pp. 369–384. doi:10.1007/978-3-030-01081-2_25.

[4] B. Kluitenberg, M. van Middelkoop, R. Diercks, H. van der Worp, What are the Differences in Injury Proportions Between Different Populations of Runners? A Systematic Review and Meta-Analysis, Sports Medicine (Auckland, N.Z.) 45 (2015) 1143–1161. doi:10.1007/s40279-015-0331-x.

[5] C. Feely, B. Caulfield, A. Lawlor, B. Smyth, Using case-based reasoning to predict marathon performance and recommend tailored training plans, in: Case-Based Reasoning Research and Development, Springer International Publishing, 2020, pp. 67–81.

[6] C. Feely, B. Caulfield, A. Lawlor, B. Smyth, Providing explainable race-time predictions and training plan recommendations to marathon runners, in: Fourteenth ACM Conference on Recommender Systems, RecSys '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 539–544. URL: https://doi.org/10.1145/3383313.3412220. doi:10.1145/3383313.3412220.

[7] C. Feely, B. Caulfield, A. Lawlor, B. Smyth, An extended case-based approach to race-time prediction for recreational marathon runners, in: Case-Based Reasoning Research and Development - 30th International Conference, ICCBR 2022, Nancy, France Septemver 12-16, 2022, Proceedings, 2022.

[8] C. Feely, B. Caulfield, A. Lawlor, B. Smyth, A case-based reasoning approach to predicting and explaining running related injuries, in: International Conference on Case-Based Reasoning, Springer, 2021, pp. 79–93.