A Better Predictor of Marathon Race Times based on Neural Networks

Dimitris C. Dracopoulos

University of Westminster, London, UK
d.dracopoulos@westminster.ac.uk

Abstract. A novel application of artificial neural networks is presented for the prediction of marathon race times based on performances in races of other distances. For many years Riegel's formula was used for the prediction of time in running races, given the race time of a person in a different distance. Recently, two different models which perform better than the classic formula in the prediction of marathon times were published in the literature. This work shows how a new approach based on artificial neural networks outperforms significantly these recently published models for marathon time prediction.

Keywords: Marathon Time Race Prediction, Neural Networks, Prediction, Running

1 Introduction

During the last decade, the sport of long distance running has become popular for amateur runners. It is common for many millions of such runners to participate in race events from 5K to a marathon each year.

Predicting how fast one can go in a future race based on performances in recent races of other distances is important for recreational runners (but also for elite athletes). Besides curiosity, the two main crucial reasons for the prediction are:

- how to design a training plan (training which takes several months before a race) based on the target race time
- what pace to follow during the race

The latter is essential not only for achieving the desired time but also a pace which is too quick in the beginning of a race can lead to "catastrophic" results, especially in longer distances such as the marathon.

2 Current Approaches for Race Time Prediction

For many years the Riegel formula [1] has been used to predict race times based on a recent race time in a different shorter distance. There are many time predictors available on the Web, e.g. the web page of the well known to runners

Runner's World magazine [2]. Most of these predictors have implemented the Riegel formula which was considered the most accurate until very recently:

$$time_{race2} = time_{race1} \cdot (\frac{distance_{race2}}{distance_{race1}})^k \quad (1)$$

where $race1$ is the shorter recent race, $race2$ is the race for which the prediction is made and k is the "fatigue factor" typically set in the range of values of $k = 1.05$ to $k = 1.07$, while in the case of world-class runners its value is set to $k = 1.08$ [1].

Recently, two new models were published by Vickers and Vertosick [3] which perform significantly better than the classic Riegel formula in the prediction of marathon race times. The formulas were adopted by *Runner's World* to implement a new predictor in the case of the marathon distance [4, 5]. The two models implementing the formulas are based on the prediction of a marathon race time either using one shorter distance race time or two shorter distance race times respectively.

In their work, Vickers and Vertosick [3], collected data for 2303 recreational endurance runners via a questionnaire published in the news website *Slate.com*. Although other studies exist for elite runners [6], the performance and race time prediction for the recreational runners have been poorly addressed [3]. After the validation of data from the questionnaires, there were only 493 runners who ran a marathon and two other races of different distances (after dropping the very fast and the very difficult races reported by the users) and these were the data used for reporting their published results for the marathon race time prediction. All the data are available to the public via [3] (additional File 2). The 493 runners data include:

- $N = 337$ data used for training, i.e. from the overall dataset in the aforementioned file, the runners who were in *group 1* or *group 2* and where *cohort3* was 1 (ran a marathon and two other races of different distances).
- $N = 156$ data used for testing, i.e. from the overall dataset in the aforementioned file, the runners from *group 3* and where *cohort3* was 1.

Some times reported by people answering the questionnaire were associated with "difficult" and "fast" races. These times were adjusted in order to be more representative of a runner's time for an "average" difficulty race of that distance. To do so, Vickers and Vertosick created a model to predict race velocity in $meters/sec$ for each race distance separately, adjusted for race difficulty (difficult, average or fast). Based on the differences in speed between average and difficult races and average and fast races, some velocity coefficients were calculated for each race distance (marathon, half marathon, 10 miles, 10K, 5 miles, 5K). These coefficients were added to the true velocity reported by each runner, in order to calculate the adjusted times (for more details of this see [3] and their additional File 1 which contains the values of the coefficients). Both original and adjusted times are included in the full dataset in the aforementioned dataset file (additional File 2).

The analysis in [3] showed that from all the collected data, the independent variables which are required to predict a longer distance race time are:

1. the time(s) of recent shorter distance race(s)
2. the typical weekly mileage while training for the longer distance race

The two models that were developed by Vickers and Vertosick for the marathon race prediction which performed significantly better than Riegel's formula were:

Model 1

Predict the marathon race time based on a single recent race of a shorter distance:

$$v_{Riegel} = \frac{42195}{time_{race1} \cdot (\frac{42195}{distance_{race1}})^{1.07}} \tag{2}$$

where $time_{race1}$ is the adjusted time for the shorter race in the case of a difficult or a fast shorter race.

$$velocity_1 = 0.16018617 + 0.83076202 \cdot v_{Riegel} + 0.06423826 \cdot \frac{typical\_mileage}{10}) \tag{3}$$

$$time_{marathon} = \frac{42195}{60 \cdot velocity_1} \tag{4}$$

where $time_{marathon}$ is the predicted time (in minutes) for the marathon race.

Model 2

Predict the marathon race time based on two recent races r_1, r_2 of shorter distances (where the distance for r_2 is longer than the distance of r_1:

$$k_{\frac{r_2}{r_1}} = \frac{\ln(\frac{time_{r_2}}{time_{r_1}})}{\ln(\frac{distance_{r_2}}{distance_{r_1}})} \tag{5}$$

where $time_{r_1}, time_{r_2}$ are the adjusted times for the shorter races in the case of a difficult or a fast shorter race.

$$k_{marathon} = 1.4510756 - 0.23797948 \cdot k_{\frac{r2}{r1}} - 0.01410023 \cdot \frac{typical\_mileage}{10} \tag{6}$$

$$time_{marathon} = \frac{time_{r_2} \cdot (\frac{42195}{distance_{r_2}})^{k_{marathon}}}{60} \tag{7}$$

where $distance_{r_2}$ is the distance of race r_2 the longer of the two shorter distance races and $time_{marathon}$ is the predicted time (in minutes) for the marathon race.

3 Neural Implementation and Results

The aim of the work described here is to develop models for the prediction of marathon race times based on neural networks and compare them with the current state of the art predictors described in the previous section.

The standard multilayer perceptron was used for the development of two different predictor models in alignment with the two Vickers-Vertosick models [3]:

- **Neuromodel 1**: Marathon time predictor given a recent race time of a shorter distance
- **Neuromodel 2**: Marathon time predictor given two recent race times in shorter distances

Two different feedforward neural networks were trained based on backpropagation, for each of the two model cases above. In both cases, the Levenberg-Marquardt method was used for the optimisation of the weights of the networks.

Both the training and the testing of the networks was done using the same datasets as the ones used in [3], in order to compare the neural networks performances directly with the Vickers-Vertosick and Riegel predictors. Thus, from the 493 data described in the previous section, the first 337 were used for training purposes and the last 156 data for testing. In the case of the neural networks approach, the 337 data were further divided into 294 data (the first of the 337) for training and 43 (the last of the 337) for validation. The 294 data are used for training which is stopped when the error on the validation set (43 data) starts increasing.

Different topologies were tried with both 1 and 2 hidden layers with a variable number of neurons in each, in order to determine the optimal topology for each of the two neuromodels. The optimum topologies reported below were decided after training all combinations of neural networks with 1 and 2 hidden layers with $5 - 20$ neurons in each layer. Thus $16 \times 16 = 256$ neural networks were trained with 2 hidden layers and 16 neural networks were trained with a single hidden layer. Each of these networks were trained separately for 10000 times, i.e. each of the 10000 times the network was initialised with different initial weights. The training time for each network took only up to a few seconds on a Quad-Core AMD Opteron 2.2GHz machine. This iterative procedure gave the optimum topologies described next.

3.1 Neuromodel 1

The neural network consisted of 3 inputs. The first two are the distance $distance_{r_1}$ and the adjusted time $time_{r_1}$ for the shorter race $r1$. The third input was the typical weekly mileage while training for the marathon race.

The optimised trained network used two hidden layers with 7 and 12 neurons respectively. The prediction was the marathon race time in minutes.

3.2 Neuromodel 2

The neural network consisted of 5 inputs. The first two are the distance $distance_{r_1}$ and the adjusted time $time_{r_1}$ for the shorter race $r1$. The third and fourth inputs are the distance $distance_{r_2}$ and the adjusted time $time_{r_2}$ for the second shorter race $r2$, where $distance_{r_2} > distance_{r_1}$. The fifth input was the typical weekly mileage while training for the marathon race.

The optimised trained network used two hidden layers with 5 and 12 neurons respectively. The prediction was the marathon race time in minutes.

3.3 Results

The same metrics that were used in [3] were calculated for the same test data (156 data), in order to have a direct comparison of the neural approach with the improved Vickers-Vertosick predictors (improved in terms of performance compared with the Riegel formula). These metrics were the mean square error and the penalised mean square error. Since overestimation of a runner's velocity is more detrimental than underestimation (a runner who starts too slow can speed up during a race whereas a runner who starts too fast will usually slow dramatically) the penalised mean squared error is calculated so that an overestimate of velocity has double the weight of an underestimate [3]. The two errors are shown below:

$$mse = \sum_{i=1}^{N} [target_{time}(i) - predicted_{time}(i)]^2 \tag{8}$$

$$\begin{aligned} penalised\ mse = & \sum_{i \in target_{time} > predicted_{time}}^{N} [2 \cdot (target_{time}(i) - predicted_{time}(i))]^2 \\ & + \sum_{i \in target_{time} \leq predicted_{time}}^{N} [target_{time}(i) - predicted_{time}(i)]^2 \end{aligned} \tag{9}$$

where $N = 156$ the size of the test data and all times are in minutes. In (9), the first summation term corresponds to the overestimate of the prediction (the time of the prediction is faster than the actual target) and thus it has the double weight of the second summation term (the predicted time is slower than the actual target).

Table 1 contains the errors for both the MSE and the penalised MSE for the three approaches, i.e. the Riegel formula, the Vickers-Vertosick (V-V) with one shorter race (model 1) and two shorter races (model 2) and the Neuromodel 1 (one shorter race) and Neuromodel 2 (two shorter races). The Riegel errors were calculated based on the longest race time available which was shorter than the marathon distance. A value of $k = 1.07$ was used for the Riegel formula. All errors are calculated for the test data $N = 156$.

Table 1. Comparison results among the Riegel formula, the two Vickers-Vertosick (V-V) models and the two Neuromodels for marathon race time prediction (Neuro 1 and Neuro 2 corresponding to Neuromodel 1 and Neuromodel 2 respectively).

	Riegel	**V-V 1-input**	**Neuro 1**	**V-V 2-inputs**	**Neuro 2**
MSE	354.7152	227.593808	172.065806	208.289713	159.457859
Penalised MSE	1318.625974	646.096737	454.432942	524.977394	394.612895

It is clear that both the Neuromodel 1 and the Neuromodel 2 perform significantly better than the two models recently introduced in [3], for the task of marathon race time prediction.

4 Conclusions

The task of long distance race time prediction based on performances in races of shorter distances is important both to recreational and elite runners. This is because the design of an individual training plan is largely based (among other factors) on the target race time and also because the pace which a runner follows during a race depends on the predicted time. Given the fact that many millions of runners participate in races every year gives extra motivation and importance to tackle this problem by making the prediction as accurate as possible.

For many years the same formula was used for this prediction task and only recently two new models were introduced which outperform the classic formula previously used for marathon time prediction.

The work here shows that approaches based on feedforward neural networks perform significantly better than these two newly introduced models.

The derived neural networks (neuromodel 1 and neuromodel 2) together with other Matlab code and files which can be used to reproduce the results of this paper can be found in: https://github.com/ddracopo/race-prediction.

Acknowledgements. The author would like to thank Andrew J. Vickers and Emily A. Vertosick for making available to the public the data collected for the derivations of their models and also for answering questions regarding their usage of the data to derive their published results.

References

1. Riegel, P.S.: 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 (1981), http://www.jstor.org/stable/27850427
2. Runner's World race time predictor, https://www.runnersworld.co.uk/health/rws-race-time-predictor
3. Vickers, A.J., Vertosick, E.A.: An empirical study of race times in recreational endurance runners. BMC Sports Science, Medicine and Rehabilitation 8(26) (2016)

4. Here's a better marathon time predictor: Your old calculator was doing it wrong. Runner's World, https://www.runnersworld.com/marathon-training/heres-a-better-marathon-time-predictor
5. Race time predictor, http://www.runnersworld.com/tools/race-time-predictor
6. Karp, J.R.: Training characteristics of qualifiers for the U.S. Olympic marathon trials. International Journal of Sports Physiology and Performance 2(1) (2007)