Predicting the Personal-Best Times of Speed Skaters Using Case-Based Reasoning

Barry Smyth[1](✉) and Martijn C. Willemsen[2]

[1] Insight Centre for Data Analytics, University College Dublin, Dublin, Ireland
Barry.Smyth@ucd.ie

[2] Jheronimus Academy of Data Science, Eindhoven University of Technology, Eindhoven, The Netherlands
M.C.Willemsen@tue.nl

Abstract. Speed skating is a form of ice skating in which the skaters race each other over a variety of standardised distances. Races take place on specialised ice-rinks and the type of track and ice conditions can have a significant impact on race-times. As race distances increase, pacing also plays an important role. In this paper we seek to extend recent work on the application of case-based reasoning to marathon-time prediction by predicting race-times for speed skaters. In particular, we propose and evaluate a number of case-based reasoning variants based on different case and feature representations to generate track-specific race predictions. We show it is possible to improve upon state-of-the-art prediction accuracy by harnessing richer case representations using shorter races and track-adjusted finish and lap-times.

Keywords: CBR for health and exercise · Speed skating · Race-time prediction · Case representation

1 Introduction

Speed skating has a long history as a popular winter sport. The International Speed Skating Union was founded in 1892 and long-track speed skating has been an Olympic sport since 1924 [1]. Olympic events include sprints (500/1000 m), middle distance (1500 m) and long distance (3,000/5,000/10,000 m) races, which impose different physiological, fitness, and pacing demands on skaters. Fast skating requires a high degree of technical skill, physical strength and dexterity: the crouched body position with low knee and body angles, which is optimal over shorter distances, is exceedingly difficult to maintain over longer distances [1,2]. Speed skating is also a time-trial event, with two skaters competing in separate

Supported by Science Foundation Ireland through the Insight Centre for Data Analytics under Grant Number 12/RC/2289P2.

I. Watson and R. Weber (Eds.): ICCBR 2020, LNAI 12311, pp. 112–126, 2020.
https://doi.org/10.1007/978-3-030-58342-2_8

lanes, so their performance mostly depends on their own abilities. Choosing a pacing strategy that is optimal, given the distance, track, competition, and the skater's own ability is a challenge and it is interesting to consider whether we can help skaters to achieve new personal-best (PB) times by recommending more appropriate pacing strategies.

There is a growing interest in the use of machine learning techniques in sports for performance prediction [3]. For example, recent research by [4–6] has considered performance prediction among marathon runners, where pacing also plays a role, showing how case-based reasoning can be used for PB prediction and pacing recommendation. Briefly, by reusing a case-base of past *race progressions*, each documenting the progress of a runner from a non-PB to a PB race, it was possible to predict challenging but achievable PBs for runners with upcoming races, based on the PBs of similar runners, and also to recommend a pacing plan to help a runner achieve their predicted PB time.

In this paper we explore whether this approach can be adapted to predict the race-times of skaters, bearing in mind that there are important and obvious differences between speed skating and marathon running. For example, speed skaters compete over a range of distances and thus there is an opportunity to create cases using multiple past races over different distances, unlike the marathon-to-marathon format of the cases used by [4–6]. This also facilitates prediction for distances that the target skater has not yet raced. While weather conditions are no doubt important in marathons, such factors were not considered by [4–6]; although a simple weather adjustment was use for ultra-running prediction by [7]. In skating the condition of the ice and the environment of the track are significant enough that they need to be included, especially since a skater's prior races will tend to take place on a variety of different tracks; we will describe how to normalise performances with respect to different tracks.

The remainder of this paper is organised as follows. In the next section we introduce speed skating as our domain of interest, discussing the important aspects of the sport, summarising the dataset that we will use, and highlighting the main research questions that we wish to answer. Following this, we will present our main technical contribution, by describing a case-based approach to predicting track-specific race-times. In fact, we will describe a number of variants of this approach, which differ in terms of the race histories that are used in cases, and the way that they are used. Finally, before concluding, we will describe a detailed evaluation to compare the prediction accuracy of these different variants, showing how significant improvements in prediction accuracy can be achieved relative to the state-of-the-art baseline approach proposed by [4,8].

2 An Introduction to Speed Skating

Speed skating is a unique sport that combines endurance and power with pacing strategy and racing aerodynamics. In this section we briefly review the major features of the sport before describing the details of the dataset used by this work. We then go on to outline the key research questions that will be considered by this research.

2.1 The Anatomy of Speed Skating: Skaters and Races

Long-track speed skating is typically performed on an 400 m artificial ice-rink (see Fig. 1) over the following distances:

1. *Sprint:* 500 m (comprising one straight end and one lap) and 1,000 m (2.5 laps);
2. *Middle Distance:* 1500 m (an opening of 300 m, with 3 additional laps) is an important distance because it combines elements of sprint and endurance skating;
3. *Endurance:* 3,000 m (7.5 laps), 5,000 m (12.5 laps) and 10 km (25 laps), all of which demand a considerable degree or pacing strategy from skaters.

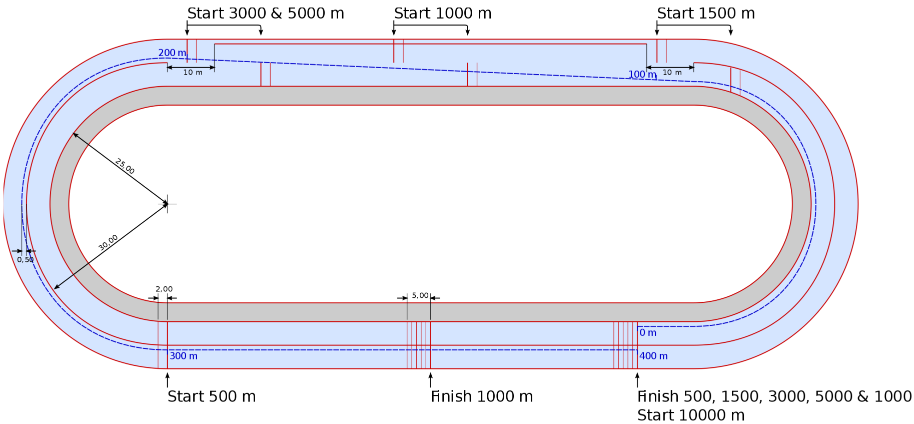

Fig. 1. The dimensions of a standard speed skating track and race configurations; image provided courtesy of wikipedia.org

In competition, skaters achieve similar high speeds to cyclists: elite sprinters reach 60 km/h while endurance skaters sustain average speeds in the 45–50 km/h range. During a race, skaters have access to very limited information on their performance – unlike runners and cyclists, GPS devices are useless as most tracks are semi-covered or completely indoor – and typically they only have access to their 400 m lap-times. Speed skating also places very different physical demands on athletes, compared with running or cycling: the crouched body position and low knee and trunk angles that are required for aerodynamic skating are physiologically challenging because they restrict blood-flow to the active muscles [1]. This makes it especially difficult for skaters to maintain good form and pacing over longer distances.

2.2 The Importance of Pacing

Previous research has focused on the pacing strategies used by elite skaters for shorter [2,9] and longer distances [10]. For sprint distances (500/1000 m), pacing does not play a significant role and the best approach is typically an *all-out* strategy with skaters going as fast as possible from the start and maintaining this speed, as best they can, until the finish [9]. However, for distances $\geq$1500 m, which combine elements of anaerobic and aerobic exertion, pacing plays an increasingly important role [11,12]. If a skater starts too fast, then they run the risk of slowing during the final stages of a race, and research has shown how maintaining a high speed in the 3rd lap (from 700–1,100 m) of 1,500 m races is critical for faster finishing times; see also the work of [13] for an analysis of a similar phenomenon among marathon runners.

Pacing is even more important in long-distance races, but in a way that differs from marathon running. For example, in elite long-distance skating negative splits – where the skater achieves a faster second-half time than first-half time – are more rare than in elite long-distance running, likely due to the physiological constraints and reduced blood-flow that is associated with good skating form.

For non-elite skaters lap-times typically slow as a race unfolds but the degree of slowing depends on the race distance: shorter races present with more significant slowdowns between laps than longer races, which are associated with more consistent pacing. As with marathons, how skaters pace their races is important when it comes to identifying similar skaters, thus motivating the importance of lap-times as part of case representations.

While previous research has focused on small samples of elite speed skaters, in this paper we focus on much larger samples of amateur and sub-elite speed skaters. Usually amateur skaters are still learning *how* to race, and thus any improvements to their pacing may enhance their PB prospects. Indeed, the pacing issue is exacerbated for non-elite skaters with respect to longer distances, in part because there are fewer opportunities to compete over longer distances, compared with elites; in other words non-elite skaters have fewer racing experiences when they need them.

2.3 The Dutch Speed Skating Data Set

The dataset used in this study was collected from http://www.osta.nl and comprises 329,080 race records from 15,590 unique Dutch skaters; thus each individual skater is associated with an average of 21 races. The races took place between September 2015 and January 2020 and race distances included all of the common distances, 500 m, 700 m, 1,000 m, 1,500 m, 3 km, 5 km, and 10 km. Each race record includes information about the skater (their name/id, gender, age), the race date, distance and track, and the skater's performance (finish-time and segment/lap-times, whether or not the result was a personal-best, and various age/gender rankings).

Skater Demographics: Speed skating is a somewhat unusual sport. In the data set skaters ranged in age from 4 to 84 years-old, but as shown in Fig. 2 most skaters are young, between the ages of 10 and 18 years-old, and once they graduate from high-school and go on to college most leave the sport, unless they

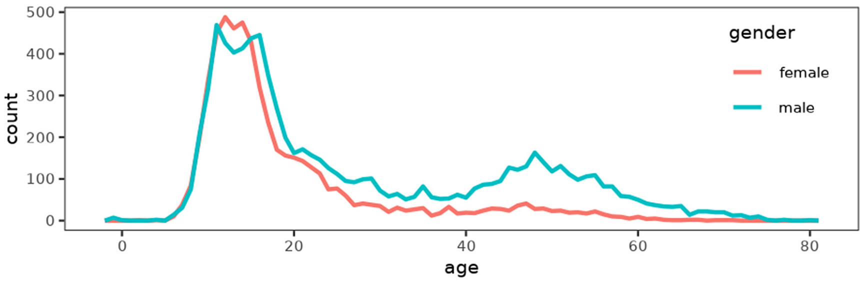

Fig. 2. Age distribution of skaters in the Dutch data set

are especially competitive. However, it is not unusual for skaters to return to the sport in their 40's, especially men, perhaps as their own children start to compete. In the Netherlands there exists a large population of older skaters who remain active at so-called *masters* level. There is even an official national masters championship.

Race Histories: There are 4 major categories of skaters in the dataset (Fig. 3: *pupils* are younger than 12 years old and only compete over shorter distances (100 m and 300 m, which are not in the data set, and 500 m, 700 m and 1000 m which are present); the majority of races are completed by *junior* skaters between 13 and 18 years old, mostly in races up to 3000 m; *senior* and *masters* skaters more frequently compete in 5 km and 10 km races, although they still remain rare compared to shorter distances, in part at least because the economics of ice-rinks make longer races more costly. The 500 m races are the most common by far because skaters often combine them with another distance on the same day or at the same event.

Track Types and Track Conditions: Track type and the ice conditions are important factors that influence performance. The quality of the ice can have an impact on race-times and is determined by a variety of factors including humidity and temperature. Outdoor or semi-covered tracks require frequent reconditioning of the ice (often every 20–30 min), while air-conditioned, closed-roof tracks provide more stable conditions, which are conducive to faster racing; high-altitude tracks are also considered to be faster [10], due to reduced air-resistance, but they are not present in the Dutch data set.

The data set contains records for a variety of track types, including: fully enclosed, air-conditioned tracks like the one in Heerenveen (HV), which hosts many international races; enclosed tracks without air-conditioning, typically with direct ventilation; semi-closed tracks with some cover, but that are otherwise exposed to the elements; and fully outdoor tracks without any cover at all. Figure 4 shows the mean 500 m race-times for a variety of different tracks and track types, and serves to highlight just how important track types are when

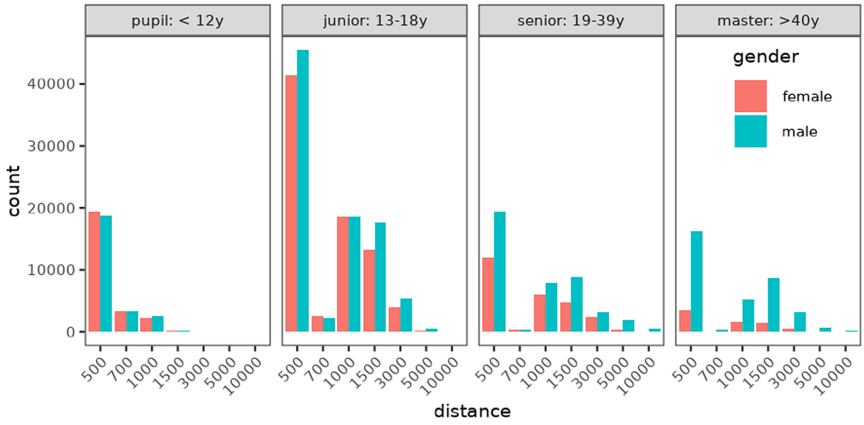

Fig. 3. Speed skaters by age category and distance.

it comes to finish-times. The fastest air-conditioned tracks are associated with finish-times that can be >10% faster that outdoor tracks (e.g. ≈45 s vs ≈53 s, for HV vs AM).

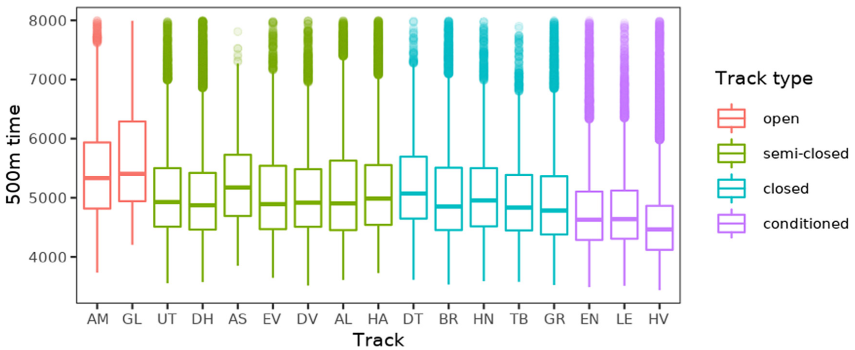

Fig. 4. Boxplots of 500 m times by track; note times are in hundreds of a second.

2.4 Research Questions

The main research question to be explored in this work is whether it is possible to accurately predict track-specific, personal-best times for skaters based on their previous racing histories. Unlike the work of [4,5,8], which relied on marathon race records from the same course, in this work each skater can be represented by a more diverse mix of race distances across a variety of tracks. As such, the main question becomes whether it is possible to predict the performances of skaters over distances that are longer than they are used to, and for different types of tracks, using their shorter racing histories. This use-case is particularly important to younger skaters because, when younger skaters graduate from shorter sprints to longer endurance races they can benefit from advice about realistic goal-times and pacing strategies.

For a given distance, differences in finish-times depend on track conditions, but they also depend on the skaters. Therefore it is not enough to simply apply a *one-size-fits-all* weighting to account for track differences when trying to predict track-specific finish-times. For example, some tracks might attract very young or much older skaters, who tend to be slower, while faster tracks like Heerenveen (HV) tend to attract more competitive skaters, who want to improve their PB, or those who wish to qualify for national championships.

3 Predicting Track-Specific Race-Times

Our approach to predicting finish-times is fundamentally case-based in nature: to predict a finish-time for some skater s and distance d we reuse the finish-times of skaters with similar race histories. To do this we describe a number of different ways to represent race histories with or without track-specific adjustments, and outline how the resulting case-bases can be used to generate predictions.

3.1 From Races to Cases

The work of [4,8] proposed pairing a runner's non-PB marathon time (and 5 km split-times) with their PB time (and split-times). The equivalent representation in the present work would, for a given target distance, d, pair a skater's non-PB race for d, $nPB(s,d)$, with their PB time for d, $PB(s,d)$ as per Eq. 1; each race, is represented by a finish-time, lap-times and a track id. In other words, to predict the finish-time for s for an upcoming 3,000 m race, requires a case-base that is made up of PB/non-PB times for 3,000 m races by other skaters. In what follows we refer to this as the *nPB* case representation (c_{nPB}) and it will serve as the *baseline* against which to judge the variations that follow.

$$c_{nPB}(s,d) = \Big\langle nPB(s,d) \mid PB(s,d) \Big\rangle \tag{1}$$

While this baseline remains valid in the present work, we are also interested in predicting a target distance PB by using previous races from shorter (more common) distances. Thus, one variation pairs a skater's PB and lap-times for shorter distances with their PB for a longer target distance, d, as in Eq. 2; in this study the target distances used are 500 m, 1,000 m, 1,500 m, 3,000 m, and 5,000 m. We refer to this as the *PB* representation (c_{PB}).

$$c_{PB}(s,d) = \Big\langle PB(s,d')\ \forall_{d'<d} \mid PB(s,d) \Big\rangle \tag{2}$$

In this way each case encodes additional performance information for s – their finish-times (and lap-times) for multiple shorter races – but these times are also personal-best times, reflecting recent *best-efforts* over these distances. This contrasts with the *nPB* representations, where it is less clear if the transition from nPB to PB is representative of a typical progression for a skater, or an artefact of the pairing of an outlier nPB with a very good PB. Moreover, the pacing patterns reflected in the lap-times of these shorter distance PBs encode important information about the type of pacing employed by a skater, which is important when it comes to finding cases that are similar in terms of their finish-times and pacing strategy: a sprinter will likely use a different pacing strategy on a 3000 m than an endurance skater, for example.

Of course, we can also combine the *nPB* and *PB* representations, so that cases for some race distance d are made up of a nPB race for that distance and PB races for shorter distances, as shown in Eq. 3, which we refer to as the *combined* representation (c_{com}).

$$c_{com}(s,d) = \left\langle PB(s,d')\ \forall_{d'<d}\ , nPB(s,d) \mid PB(s,d) \right\rangle \tag{3}$$

3.2 Adapting for Track Variations

Given that track conditions can have a material impact on finish-times we also produce modified versions of the above case representations, which use adjusted finish-times to reflect these conditions. In our initial analyses we found that simple adjustments for mean times per track (as reflected in Fig. 4) did not improve our predictions, because there are many confounding factors at play, such as different track-specific populations and type of races.

Since many skaters in our dataset have race times for a specific distance, on different tracks, we can estimate within-person adaptations that overcome most of these confounds. For each skater and each distance we calculate a PB for each track they have raced on, and then fit a multilevel regression model to this data to estimate within-person, track-specific differences relative to a single reference track. The fixed effects of this multilevel regression model provide the adjustments that can be used to standardise the finish-times of all races relative to the reference track.

$$c'_{nPB}(s,d) = \left\langle nPB'(s,d) \mid PB'(s,d) \right\rangle \tag{4}$$

$$c'_{PB}(s,d) = \left\langle PB'(s,d')\ \forall_{d'<d}\ \mid PB'(s,d) \right\rangle \tag{5}$$

$$c'_{com}(s,d) = \left\langle PB'(s,d')\ \forall_{d'<d}\ , nPB'(s,d) \mid PB'(s,d) \right\rangle \tag{6}$$

These adjusted finish-times can then be used to produce new versions of our *nPB, PB* and *combined* case representations, as shown in Eqs. 4–6, by replacing *raw* timing data with *normalised*, track-adjusted timing data, as indicated by nPB' and PB'.

3.3 Generating Predictions

For a given skater s and race distance d we predict their finish-time (using one of the representations outlined above) by using the past races of s to identify the k nearest cases, using a standard Euclidean distance similarity metric. As in the work of [4,5,8], male and female skaters are separated so that the predictions for male skaters are generated from the cases of male skaters, and vice versa for females; this is because of the performance differences that exist between the sexes due mainly to physiological differences. We also separate younger skaters (≤20 years-old) from older skaters (≥40 years-old) to facilitate a later age-based comparison.

Then a prediction is generated based on the distance-weighted mean of the target distance PB times from these cases ($PB(s', d)$ or $PB'(s', d)$ as appropriate, where s' denotes a similar, nearest-neighbour skater). If adjusted timings are used then ($PB'(s', d)$) then obviously the resulting prediction needs to be transformed back into an actual finish-time for the target track. As an aside, it is worth noting that to predict a pacing plan for the target race we can adopt a similar approach to that described in [4,8], by computing the average relative lap-times from the k nearest cases. However, we do not focus on this particular task further in this paper.

4 Evaluation

In this section we provide a detailed analysis of prediction accuracy, comparing the baseline *nPB* approach originally described in [4,8] to the alternatives proposed in this work.

4.1 Data and Methodology

We use the Dutch dataset introduced earlier to produce different case-bases for three common target distances (1,500 m, 3,000 m, and 5,000 m), using the different case representations (nPB, PB, and combined), and timing data (raw times versus adjusted times). This leads to 18 (3 × 3 × 2) individual case-bases for prediction. Note that the different target distances have quite different race characteristics: there are ≈48k 1,500 m races, each with 4 lap times, compared with ≈16k 3,000 m races (each with 8 lap times) and ≈2.7k 5,000 m races (with 13 lap times per race). The longer distances also facilitate richer *PB* representations because there are more shorter component PB distances. Thus a 5,000 m *PB* or *combined* case will have significantly more features than a 5,000 m *nPB* case, because of its extra component PB cases, and their lap-times.

We adopt a standard 10-fold cross-validation approach to evaluate prediction accuracy across these variations and for different values of k (1, 3, 5, 10. 20, 50). During each fold/iteration we select a random 10% of cases to use as test problems with the remaining 90% of cases used as the training case-base. Each test problem is solved (generating a race-time prediction) and compared to the known race-time for that test problem. For each prediction we calculate a percentage error and compute an average error across the folds for each variation.

4.2 Prediction Error vs *k*

To begin, it is informative to explore how prediction error varies with k, the number cases retrieved to make a prediction, and how this depends on the target distance, representation, and whether or not track-adjusted timings are used. Figure 5 shows the results, separately for each combination of (a) target distance, (b) representation, and (c) timings.

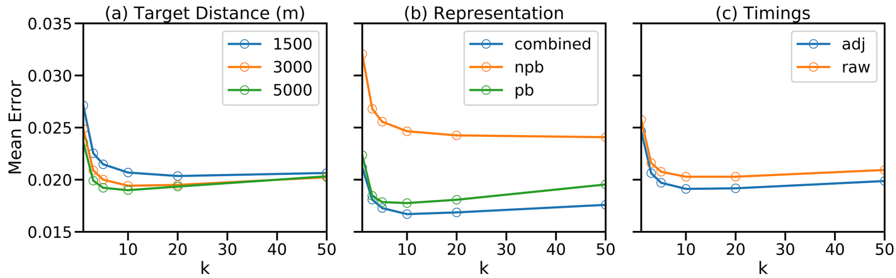

Fig. 5. The mean error rates by k (the number of similar cases reused) for different (a) target distances, (b) representations, and (c) timings (raw versus track-adjusted).

In general, as we might expect, the accuracy of predictions improves with k, up to a point, and on average the best overall errors are available for values of k in the range 10–20. It is also clear that the accuracy of the predictions, for a given value of k, depends on the target distance, representation and timing and it is worth discussing these accuracy differences further before proceeding.

The different error rates between the target distances can be explained by the number and quality of the features used during prediction. Since the error rates in Fig. 5(a) are averaged over all representations and timings, then cases for longer target distances tend to use more features, since the *combined* and *PB* representations will be made up of additional PB races and because longer races will be made up of more lap-time features. This explains the lower *combined* and *PB* error rates in Fig. 5(a). Moreover, since these longer distances are predominantly skated by the more skilled skaters, they are more predictable even at lower k. However, for larger k, the error goes up, most likely because there are fewer records for the longer distances and a larger k results in less representative similar cases being reused. This does not apply to the 1500 m distance, which still benefits from larger k due to the much larger number of available race records, and good similar cases can still be found even up to $k = 50$.

A related argument can be applied to explain the error differences by representation, in Fig. 5(b): the *combined* cases contain more features than the *PB* cases which, in turn, contain more features than the *nPB* cases. In addition, it is reasonable to expect that the PB races used in the *PB* and *combined* cases will be of *higher quality*, from a prediction viewpoint, than the lone nPB races in the *nPB* cases. That being said, the improved error rates for the *combined* cases over the *PB* cases indicates that these nPB races do add still value.

Figure 5(c) presents a like-for-like comparison in terms of the number and types of features in cases, the only difference being whether raw timings or track-adjusted times are used. The difference in error is more modest across values of k but indicates a benefit accruing to the track-adjusted timings.

In summary then, the novel case representations (*PB* and *combined*) and track-adjusted timings proposed in this work lead to more accurate predictions than the baseline *nPB* representation from [4,8].

4.3 Best Performers

Given the sensitivity of prediction to k, the target distance, the case representation, and the timings used, it is appropriate to examine the single best performing k for each combination of distance, representation, and timings, so that we can compare individual systems (single case-based predictors) more directly.

Figure 6 presents a table of these *best performers* for each of the 18 unique combinations of distance, representation, and timings. Each row of the table represents a single case-based predictor, with its corresponding value of k, and shows the mean and standard error of the prediction errors produced by the 10-fold cross-validation. The table is arranged in blocks by target distance (1,500 m, 3,000 m, and 5,000 m) and within each block the *baseline* and *best* performing variants are indicated.

We can see that best predictors, for a given combination of distance, representation, and timing, produce their most accurate predictions for different values of k, from 3 to 50, although in most cases the best value of k is either 10 or 20. The *combined* representation using track-adjusted times provides the most accurate predictions, regardless of target distance, with significant improvements with respect to the baseline, as shown. For example, when predicting 1,500 m times, the *combined*, *track-adjusted* variant generates predictions with a mean error of 0.0154 and a standard error of 0.0015, as compared with 0.0298 and .0016 for the baseline; a relative error improvement of more than 48% due to the *combined*, *track-adjusted* approach. As the target distances increase the improvements for the *combined*, *track-adjusted* variant, relative to the baseline, decrease, but remain significant; we observe a relative error improvement of 29% and 21% for 3,000 m and 5,000 m races, respectively.

It is interesting to note that these results appear somewhat at odds with the average prediction errors by target distance from Fig. 5(a), where shorter distances were associated with larger errors. While this is true in general – Fig. 5(a) averages over all representations and times for a given distance – the much higher error for the *nPB* cases for the 1500 m tends to increases the overall error rate. When we compare the *single, best performing* system for each distance, then the shorter distances have lower best-errors. This may be due to the fact that there are many more 1,500 m cases to choose from than there are for the 3,000 m or 5,000 m distances, as previously discussed.

d	rep	t	k	mean	SEM	
1500	combined	adj	20	0.0154	0.0015	*Best*
1500	combined	raw	20	0.0156	0.0016	
1500	pb	adj	20	0.0161	0.0032	
1500	pb	raw	10	0.0168	0.0034	
1500	npb	adj	50	0.0279	0.0016	
1500	npb	raw	50	0.0298	0.0016	*Baseline*
3000	combined	adj	10	0.0171	0.0017	*Best*
3000	combined	raw	10	0.0177	0.0018	
3000	pb	adj	10	0.0172	0.0024	
3000	pb	raw	10	0.0178	0.0027	
3000	npb	adj	20	0.0219	0.0017	
3000	npb	raw	50	0.0240	0.0018	*Baseline*
5000	combined	adj	10	0.0166	0.0003	*Best*
5000	combined	raw	10	0.0173	0.0003	
5000	pb	adj	5	0.0179	0.0003	
5000	pb	raw	3	0.0197	0.0003	
5000	npb	adj	10	0.0196	0.0003	
5000	npb	raw	50	0.0211	0.0003	*Baseline*

Fig. 6. Mean and standard error of prediction errors for the best performing value of k for each of the 18 case-base variants.

Figure 6 also indicates that the PB representation is also associated with significantly lower errors than nPB; the latter has fewer, lower quality features than the former. Moreover, for any given combination of distance and representation, the best track-adjusted timing cases offer improved errors compared to the use of raw timings; although the difference for a given distance and representation tends to be modest and is not commonly statistically significant.

4.4 On Gender and Age

The work of [4,8] highlighted different marathon-time prediction errors for men versus women: women enjoyed superior prediction accuracy, a result that is consistent with the notion that female runners tend to pace their marathons more evenly than men, and therefore are more predictable in their finish-times; see [14]. We consider a similar question here, by examining male and female prediction accuracy, and also the accuracy associated with younger ($\leq$20) and older ($\geq$40) skaters. We do this for two approaches – the *best* overall approach (*combined* representation with *track-based timing adjustments*) and the *baseline* (nPB with raw timings) – for the three target distances (1,500 m, 3,000 m, and 5,000 m).

We define a *relative advantage* score for gender and for age as shown in Eq. 7, so that the relative advantage for males versus females, for the baseline, is one minus the baseline error rate for males divided by the baseline error rate for females; thus, if $RelAdv_{baseline}(males, females) < 0$, then it means that the baseline error rate for males is *higher* (worse) than the baseline error rate for females, and vice versa.

$$RelAdv_{alg}(x, y) = 1 - \frac{error(x)}{error(y)} \tag{7}$$

Figure 7 presents the scores for the *best* and *baseline* approaches, for the target distances, comparing error rates for gender and age. $RelAdv_{best}(males, females) > 0$ in Fig. 7(a) means that the *best* approach produces more accurate predictions for men than for women. But $RelAdv_{baseline}(males, females) < 0$, indicating that the baseline produces more accurate predictions for women than for men, as with [4,8] for marathons. A similar pattern is observed in Fig. 7(b), comparing younger and older skaters: For the *best* approach the race-times of younger skaters are predicted more accurately $RelAdv_{best}(younger, older) > 0$ than older skaters, but for the *baseline* approach the finish-times of older skaters are predicted more accurately, $RelAdv_{baseline}(younger, older) < 0$ (except in the case of the 5,000 m target distance).

It is not clear why these approaches perform in this way, but it indicates that the *best* approach offers a more balanced prediction accuracy than the *baseline* approach, as well as its better overall accuracy. For example, the mean absolute relative advantage of the *best* approach is ≈0.05, for gender and age, indicating that the mean errors between genders and ages differ by only about 5%. This is compared with corresponding scores of 0.08 and 0.19 for *baseline*, indicating a much greater imbalance between genders and between age categories.

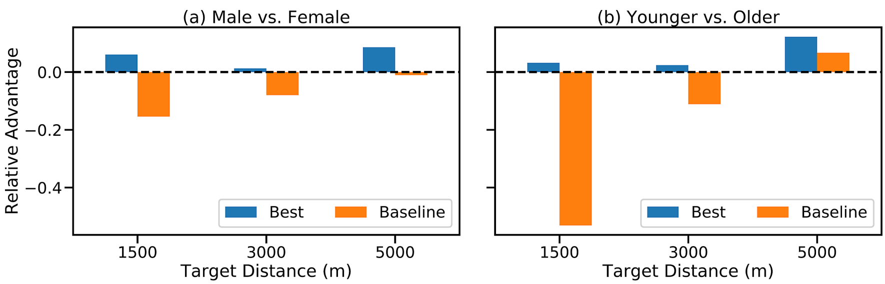

Fig. 7. A comparison of the relative error rates for the *best* and *baseline* approaches with respect to gender and age. A *relative advantage* score <0 for gender means females enjoy more accurate predictions than men, using a given approach, and a similar score for age means older skaters enjoy better predictions than younger skaters.

5 Conclusions

This paper extends the original work of [4,8], on predicting finish-times for marathon runners, in a number of important ways. First, we apply the techniques described by [4,8] to the very different sport of speed skating. Second, we propose an alternative case representation which is better suited to speed skating by representing case uses multiple races that are shorter than the target race; this in turn addresses one of the key shortcomings of the [4,8] approach, which required runners to have run at least one marathon in the past. Finally, given the importance of track conditions in speed skating we also proposed a technique for normalising race-times across a wide range of tracks. The results of a large-scale evaluation demonstrate the benefits of the new approaches that have been proposed. Using these approaches it has been possible to significantly reduce the prediction error compared with the baseline approach of [4,8].

The ideas presented in this work are general enough that they may also be applicable to marathons and other sports. For example, in marathon running, course conditions may have a significant impact on performance, which speaks to the value of a similar timing adjustment for marathon races to the one presented here for speed skating. Moreover, since many marathoners will run shorter races too (5k's, 10k's, half-marathons), then the idea of including PBs over shorter distances is also likely to be worthwhile. We also plan to extend our current work to include pacing recommendations as was the case for marathon races [5,6] to help skaters to achieve their predicted PB times and even help skaters to tackle a first race over a new, longer distance.

References

1. Konings, M.J., Elferink-Gemser, M.T., Stoter, I.K., van der Meer, D., Otten, E., Hettinga, F.J.: Performance characteristics of long-track speed skaters: a literature review. Sports Med. **45**, 505–516 (2015). https://doi.org/10.1007/s40279-014-0298-z
2. Hettinga, F.J., De Koning, J.J., Schmidt, L.J.I., Wind, N.A.C., MacIntosh, B.R., Foster, C.: Optimal pacing strategy: from theoretical modelling to reality in 1500-m speed skating. Br. J. Sports Med. **45**, 30–35 (2011)
3. Bartolucci, F., Murphy, T.B.: A finite mixture latent trajectory model for modeling ultrarunners' behavior in a 24-hour race. J. Quant. Anal. Sports **11**(4), 193–203 (2015)
4. Smyth, B., Cunningham, P.: Running with cases: a CBR approach to running your best marathon. In: Aha, D.W., Lieber, J. (eds.) ICCBR 2017. LNCS (LNAI), vol. 10339, pp. 360–374. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61030-6_25
5. Smyth, B., Cunningham, P.: An analysis of case representations for marathon race prediction and planning. In: Cox, M.T., Funk, P., Begum, S. (eds.) ICCBR 2018. LNCS (LNAI), vol. 11156, pp. 369–384. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01081-2_25

6. Smyth, B., Cunningham, P.: A novel recommender system for helping marathoners to achieve a new personal-best. In: Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys 2017, Como, Italy, 27–31 August 2017, pp. 116–120 (2017)
7. McConnell, C., Smyth, B.: Going further with cases: using case-based reasoning to recommend pacing strategies for ultra-marathon runners. In: Bach, K., Marling, C. (eds.) ICCBR 2019. LNCS (LNAI), vol. 11680, pp. 358–372. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-29249-2_24
8. Smyth, B., Cunningham, P.: Marathon race planning: a case-based reasoning approach. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, 13–19 July 2018, pp. 5364–5368 (2018)
9. Muehlbauer, T., Schindler, C., Panzer, S.: Pacing and sprint performance in speed skating during a competitive season. Int. J. Sports Physiol. Perform. **5**, 165–176 (2010)
10. Muehlbauer, T., Panzer, S., Schindler, C.: Pacing pattern and speed skating performance in competitive long-distance events. J. Strength Cond. Res. **24**, 114–119 (2010)
11. Muehlbauer, T., Schindler, C., Panzer, S.: Pacing and performance in competitive middle-distance speed skating. Res. Q. Exerc. Sport **81**, 1–6 (2010)
12. Stoter, I.K., MacIntosh, B.R., Fletcher, J.R., Pootz, S., Zijdewind, I., Hettinga, F.J.: Pacing strategy, muscle fatigue, and technique in 1500-m speed-skating and cycling time trials. Int. J. Sports Physiol. Perform. **11**, 337–343 (2016)
13. Smyth, B.: Fast starters and slow finishers: a large-scale data analysis of pacing at the beginning and end of the marathon for recreational runners. J. Sports Anal. **4**(3), 229–242 (2018)
14. Trubee, N.W.: The effects of age, sex, heat stress, and finish time on pacing in the marathon. Ph.D. thesis, University of Dayton (2011)