Constructing and Exploring a Predictive Model for 100-Meter Sprint Segmented Data

Yunhao Cui

School of Sports Engineering, Beijing Sport University, Beijing, 100084, China
2021011818@bsu.edu.cn

Abstract In a 100-meter race, segmented data every 10 meters is crucial for studying an athlete's performance. Although there are currently two ways - official provision and video analysis - to obtain segmented data, both have significant drawbacks, making it very difficult to obtain segmented data at present. To address this issue, this study created a relevant dataset, defined the corresponding input sequences, and constructed a satisfactory prediction model based on Random Forests to predict these segmented data. Users only need to input data at any position and in any quantity within the sequence to obtain accurate segmented data every 10 meters. In addition, this paper also explored the model to a certain extent, and summarized the characteristics of input sequences that contribute to the generation of high-quality prediction results: the quantity of known data should be as large as possible; the distribution of known data should be as dispersed as possible; the positions of known data should be as close as possible to each 10-meter; and the positions of known data within the first 50 meters are more favorable than those in the latter 50 meters. Under these guidelines, users of the model can better utilize it to obtain satisfactory prediction results.

Keywords: 100-meter Sprint, Segmented Data, Random Forest, Sequence Characteristics

1 Introduction

In a 100-meter race, the segmented data of athletes every 10 meters is crucial for studying the performance of 100-meter sprinters [1]. Analysts can obtain the speed variations of athletes during the race through these segmented data, thereby further analyzing the technical characteristics and issues of the athletes [2][3]. However, obtaining complete segmented data for every 10 meters is quite difficult, which results in a scarcity of existing data.

There are currently two main measurement methods, both of which have significant drawbacks. The first method involves direct on-site measurements at the competition venue. However, due to the lack of necessary equipment and personnel at the majority of current venues, conducting these measurements would require procuring a large quantity of relevant measuring devices or training numerous personnel, which makes widespread implementation quite difficult. Throughout history, in most 100-meter

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_18

races, officials have rarely provided segmented data for every 10 meters. Even in top-level events such as the Olympics and World Championships, only partial 10-meter segmented data is occasionally provided (Table 1 presents segmented data from the 2009 Berlin World Championships provided by the IAAF) [4]. Besides, it's also impossible to obtain segmented data from past races using this method since it's impossible to go back to the past to deploy the measuring devices.

Table 1 Segmented Data from the 2009 Berlin World Championships provided by the IAAF

Name	Round	t-20m	t-40m	t-60m	t-80m	t-100m
Usain Bolt	Final	2.88	4.64	6.31	7.92	9.58
	SF	2.89	4.68	6.41	8.11	9.89
	Ht	2.93	4.73	6.47	8.20	10.03
	Ht	2.94	4.77	6.55	8.33	10.20
Tyson Gay	Final	2.92	4.70	6.39	8.02	9.71
	SF	2.99	4.80	6.54	8.21	9.93
	Ht	2.97	4.79	6.53	8.22	9.98
	Ht	3.02	4.85	6.62	8.35	10.16
Asafa Powell	Final	2.91	4.71	6.42	8.10	9.84
	SF	2.92	4.73	6.47	8.17	9.95
	Ht	2.89	4.69	6.41	8.10	9.95
	Ht	2.90	4.72	6.56	8.44	10.38

The second method involves obtaining segmented data by measuring from race videos. The principle of this method is to analyze the video frame by frame through the landmark lines on the track, and then infer the data of athletes at each 10 meter segment [5] (Although there are no markings specifically indicating the positions of every 10 meters on the track, there are many other landmark lines in the running area of the 100-meter race track (Figure 1), including the 10 hurdle lines for men's hurdles, the 10 hurdle lines for women's hurdles, the starting line for the one mile run, the boundary line of the 4x100-meter relay exchange zones, and the last two hurdle lines for the 400-meter hurdles [6]. Theoretically, analysts can utilize a total of 24 landmark lines from the video to obtain data of athletes). However, in reality, not every track has all these landmark lines. Analysts usually only have access to some of these markings, which greatly increases the difficulty of our inference. Besides, this method is also influenced by subjective factors, as different individuals may have different analysis results from the same video.

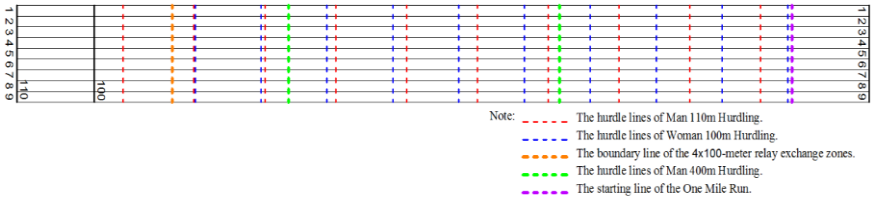

Figure 1 Landmark Lines in the Area of the 100-meter Race Track

Above all, the aim of this research is to find a method to address these current drawbacks, reduce the difficulty of obtaining segmented data, and obtain more objective and accurate analysis results. This paper regards the data of athletes at all landmark lines as a sequence. A prediction model based on random forest algorithm of machine learning will be trained by generating any number of missing values from the complete sequence. Through this model, analysts can predict the segmented data of every 10 meters of an athlete by inputting any number of the data in the sequence. Furthermore, this article will also do some exploration on the model, discuss what kind of input sequence would be more conducive to the model generating better prediction results.

2 Method

In this research, the author first defined the form of the sequence and created the corresponding raw dataset. Then the author preprocessed the dataset to construct an input-output training set suitable for model training. Then, the author trained the training set based on RF (random forest), obtained a prediction model. Finally, the author explored the characteristics of ideal input sequences when using the model for prediction. The author discussed the impact of the quantity of known data in the input sequence on accuracy during the prediction phase, as well as which positions and distributions of data within the input sequence are conducive to achieving better prediction results. Figure 2 shows the workflow of the research in this paper.

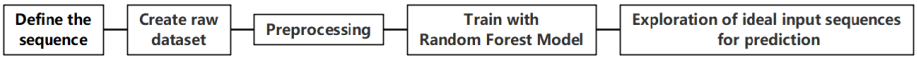

Figure 2 Research Workflow

2.1 Definition of the sequence and raw dataset creating

In the first place, the author defined the sequence that is going to be processed. The length of the sequence is 30, which includes the data at every 10 meters, and the data on the 24 landmark lines on the track mentioned in the introduction. This is because in practice, any of these 30 data points could potentially be obtained by analysts (the data on certain landmark lines measured through video, and the data at certain 10-meter intervals possibly provided by official sources such as the Olympics or World Championships). Based on this definition, the raw dataset was created (Figure 3), which was obtained by analysts through years of measurement.

3.72	10	12.86	13	20	21.5	22	25	30	31.14	38.5	40	40.28	47	49.42
0.773	1.88	2.236	2.252	2.88	3.009	3.053	3.323	3.78	3.882	4.514	4.64	4.664	5.228	5.431
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

50	55.5	58.56	60	64	67.7	70	72.5	76.84	80	81	85.98	89.5	90	100
5.48	5.94	6.192	6.31	6.632	6.926	7.11	7.311	7.663	7.92	8.001	8.408	8.699	8.74	9.58
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

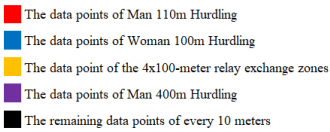

Figure 3 The Raw Dataset

2.2 Preprocessing

The aim of the trained model is to predict the segmented data of every 10 meters of an athlete by inputting any number of the data in the sequence. Therefore, the training data for the model must contain missing values of any quantity or position, enabling the model to handle arbitrary missing sequence inputs. The author used a loop to generate all possible quantities of missing values in the input sequence. Then, within the loop, the author used a random function to generate random missing positions corresponding to the quantity. Since the last column of the sequence is the athlete's competition result, which is always available, it was not included in the range for generating missing values. Next, the sequence containing missing values were used as input sequences X, and the segmented data for every 10 meters was extracted as the prediction results y. Table 2 illustrates the preprocessing process.

Table 2 Preprocessing training dataset

Algorithm 1 Preprocessing training dataset

Input:
samples: samples from the raw dataset
Output:
X: Input sequences for training
y: Output prediction results for training

```
for each sample in samples do
    for missing number from 1 to 29 do
        Randomly select non-repeating missing positions with the quantity of missing number
        Set the values at the missing positions of the sample to -1
        Append the sample with missing values to X
        Select the segmented data in every 10 meters from the sample
        Append the segmented data to y
    end for
end for
return X and y
```

Finally, apply the aforementioned process to 80% of the data from the raw dataset. Integrate all X and y as the training dataset for the model, and the other 20% data from the raw dataset was kept as the testing dataset.

2.3 Training with Random Forest Model

Random Forest (RF) is an ensemble learning method. It predicts by constructing multiple decision trees and combining them together. From a technical standpoint, its concept can be represented by the following formula:

$$Imp(X_j) = \frac{1}{M}\sum_{m=1}^{M}\sum_{t}\frac{p(t)}{p(t_{total})}\cdot imp(X_j, t) \quad (1)$$

where $Imp(X_j)$ represents the importance of variable X_j, M is the total number of trees in the random forest, t is the node index in tree m, $p(t)$ is the proportion of samples contained in node t, and $imp(X_j, t)$ is typically defined as the decrease in node impurity [7].

Random Forest improves the prediction accuracy of individual decision trees by majority voting or averaging, and demonstrates high accuracy in handling high-dimensional data and sequences. In addition, it can enhance model diversity through random feature selection and bootstrapping sampling. By averaging multiple trees, it reduces the correlation between individual decision trees, thereby reducing the likelihood of overfitting [8]. Therefore, the author chose the Random Forest model for predicting sequences.

The author trained the random forest model using X and y obtained from preprocessing, and got a model. Through this model, analysts can predict an athlete's segmented data of every 10 meters by inputting any number and any positions of the data in the sequence. And the author evaluated the model's prediction performance by the Mean Squared Error (MSE) between predicted and actual values [9]. The training results will be presented in the Result section.

2.4 Exploration of Ideal Input Sequences for Prediction

The quantity of known data in the input sequence. First, the author explored the impact of the quantity of known data in the input sequence on generating ideal prediction results, as shown in table 3.

Table 3 Exploration of the quantity of known data

Algorithm 2 Exploration of the quantity of known data

Input:
samples: samples from the testing dataset
Output:
MSE: Mean squared error of prediction under different known number

```
1   for known number from 1 to 29 do
2     for each sample in samples do
```

```
        Randomly select non-repeating known positions with the quantity of known number
        Keep only the values at the known positions of the sample and set the rest to -1
        Append the sample with known values to X
        Select the segmented data in every 10 meters from the sample
        Append the segmented data to y
    end for
    Inputting X into the trained random forest model to obtain predicted results y'
    Calculate the mean squared error between y and y', and store it in an array MSE
end for
Plot the change of MSE with the known number
```

The author varied the quantity of known data in the input sequence, and observed the changes in MSE. The output results will be presented in the Result section.

The positions of known data in the input sequence. Next, the author explored the importance of data at different positions in the sequence for generating ideal prediction results, as shown in table 4.

Table 4 Exploration of the positions of known data

Algorithm 3 Exploration of the positions of known data

Input:
samples: samples from the testing dataset
Output:
MSE list: The length of the list is equal to the length of the input sequence (29), where each position's value represents the sum of the predicted MSEs for all combinations involving that position in the input sequence.

```
for known number from 1 to 29 do
    Generate all combinations of positions where known data is located under the known number
    for each combination in combinations do
        for each sample in samples do
            Keep the values at the positions in combination of the sample and set the rest to -1
            Append the sample with values at the positions in combination to X
            Select the segmented data in every 10 meters from the sample
            Append the segmented data to y
        end for
        Inputting X into the trained random forest model to obtain predicted results y'
        Calculate the MSE between y and y'
        Add the MSE value to the corresponding combination position in the MSE list
    end for
end for
Normalize and plot the MSE list
```

The author traversed all input sequences and accumulated the MSE values for each case to the corresponding positions in the MSE list. Since each position is accumulated the same number of times, the smaller the accumulated value, the more important the position is for generating ideal prediction results. The output results will be presented in the Result section.

The distribution of known data in the input sequence. Finally, the author explored how the distribution of known data in the sequence is more favorable for the generation of ideal prediction results (e.g., discrete distribution, concentrated distribution.). The author defined an indicator - dMAE for evaluating distribution:

$$dMAE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{28(i-1)}{n-1} + 1 - P_i\right) \tag{2}$$

where n represents the number of known data in the input sequence, P_i represents the position index of the i-th known data in the input sequence. dMAE represents the dispersion degree of the known data distribution in the input sequence, where a smaller value indicates a greater dispersion. Table 5 is the exploration of the distribution of known data.

Table 5 Illustrates the exploration process:

Algorithm 4 Exploration of the distribution of known data

Input:
samples: samples from the testing dataset
Output:
dMAE-MSE_data: The data pair consisting of the dMAE of the input sequence and the MSE of the prediction result

```
for known number from 1 to 29 do
    Generate all combinations of positions where known data is located under the known number
    for each combination in combinations do
        for each sample in samples do
            Keep the values at the positions in combination of the sample and set the rest to -1
            Append the sample with values at the positions in combination to X
            Select the segmented data in every 10 meters from the sample
            Append the segmented data to y
        end for
        Inputting X into the trained random forest model to obtain predicted results y'
        Calculate the dMAE of X and the MSE between y and y'
        Store dMAE and MSE as a dMAE-MSE_data
    end for
end for
Plot all dMAE-MSE_data
```

The author traversed all input sequences, using the dMAE value of the input sequence as the horizontal axis and the MSE value of the prediction result as the vertical axis. After outputting, the relationship image between the prediction effect and the distribution of known data can be obtained. The output results will be presented in the Result section.

3 Result

3.1 Training Result of the Prediction Model

As shown in Figure 4, analysts can predict the segmented data at every 10 meters by inputting any number and any positions of data in the sequence to the model. The model will also show the MSE value to evaluate the model's prediction performance.

```
input_sequence:
 [ 0.778 -1.    -1.     2.283  2.94  -1.    -1.    -1.    -1.    -1.
 -1.    -1.     4.795 -1.    -1.     5.64   6.114 -1.    -1.     6.842
  7.161 -1.     7.579 -1.    -1.     8.328 -1.    -1.    -1.    10.03 ]
predicted_sequence:
 [1.9, 2.94, 3.87, 4.76, 5.64, 6.5, 7.36, 8.23, 9.12, 10.03]
true_sequence:
 [ 1.9   2.94  3.87  4.77  5.64  6.5   7.36  8.24  9.12 10.03]
mse:
 1.9999999999999148e-05
```

Figure 4 Demonstration of the model

The model's prediction accuracy is quite high, with MSE values consistently within 0.002, and in the vast majority of cases, it is below 0.0005. Therefore, it can be concluded that the model can effectively achieve the purpose of predicting the segmented data of athletes every 10 meters.

3.2 Results of the Exploration of Ideal Input Sequences for Prediction

The quantity of known data in the input sequence. As shown in Figure 5, as the quantity of known data increases, the MSE value becomes smaller.

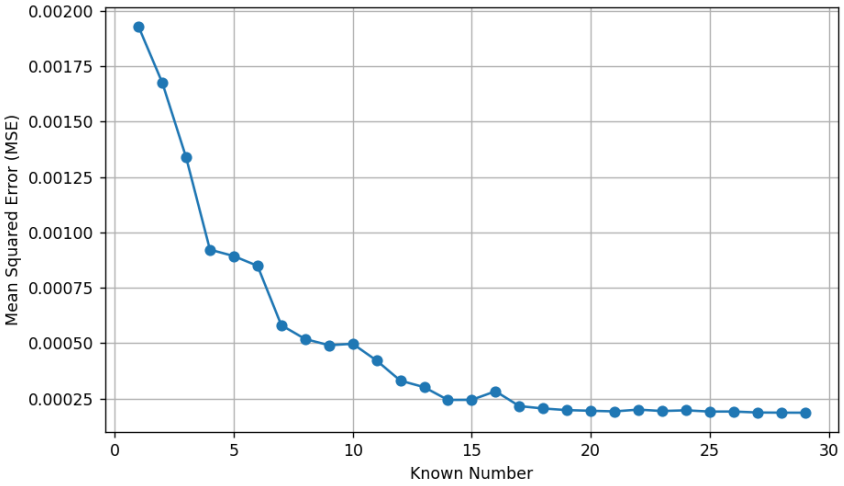

Figure 5 Relationship diagram between known number and MSE

Therefore, it can be concluded that the more known data in the input sequence, the better it is for the model to generate good prediction results.

The positions of known data in the input sequence. After normalizing the accumulated results, the importance of each data position is shown in Figure 6.

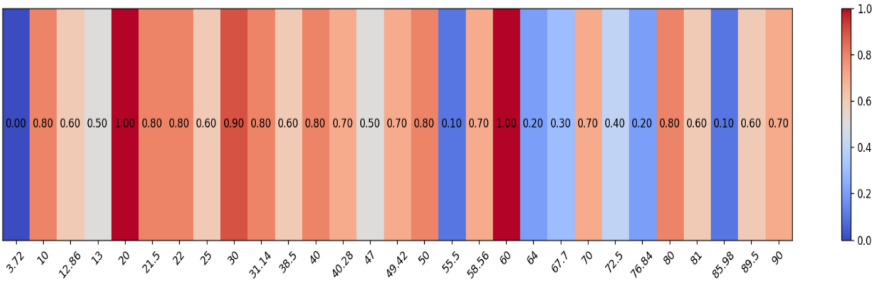

Figure 6 The importance of each data position

From this, it can be inferred that the data at the positions of each 10 meter is the most important for generating good prediction results, followed by the positions near each 10 meter location. Overall, the data positions within the first 50 meters are more important than those in the latter 50 meters, which is more conducive to the model generating good prediction results.

The distribution of known data in the input sequence. The scatter plot of dMAE values of the input sequence and corresponding MSE values of the prediction results is shown in Figure 7:

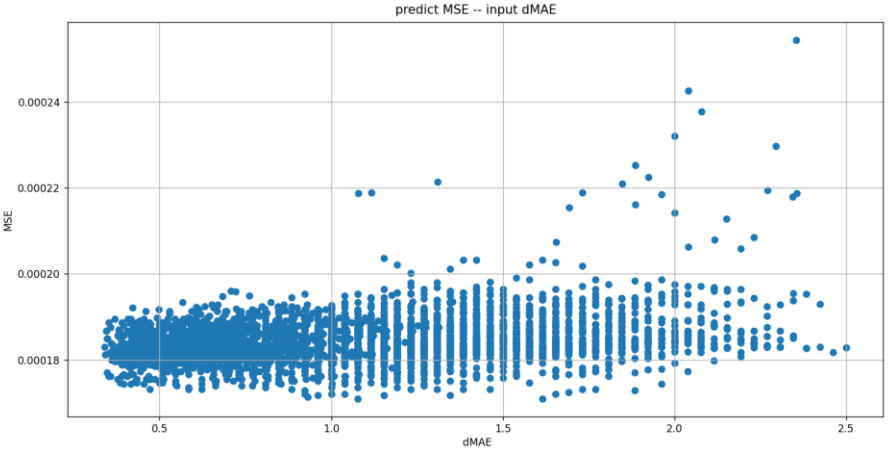

Figure 7 Scatter plot of the relationship between MSE and dMAE

It can be observed that as the dMAE value of the input sequence increases, the MSE value of the prediction results also tends to become larger. From this, it can be inferred

that the more dispersed the distribution of known data in the input sequence, the better it is to the model generating good prediction results.

3.3 Discussion

In terms of model construction, the predictive performance of the model constructed in this article is very good. This indicates that the construction of the dataset in this paper is scientific, and the preprocessing of the dataset as well as the selection of the model (random forest model) are also commendable.

In terms of model exploration, this article explored the ideal input sequence of the model from three perspectives: the quantity of known data in the input sequence, the positions of known data in the input sequence, the distribution of known data in the input sequence. After exploration, the following conclusions were drawn: the more known data in the input sequence and the more dispersed the distribution, the more conducive it is for the model to generate a good prediction result. In addition, when using the model, data should be input as much as possible for every 10 meters and nearby positions, and it is advisable to input data from positions within the first 50 meters as much as possible. This will be more conducive to the model generating better prediction results. At the same time, this article also has certain limitations. Due to the small size of the dataset used in this article, and most of them are results of elite athletes within 10 seconds, the model may overfit within the range of elite athletes and may not be suitable for athletes at other levels, resulting in poor generalization [10].

4 Conclusion

In this study, a model was constructed to predict segmented data for the 100-meter race and explored to a significant extent. In the construction phase, the author initially compiled a relevant raw dataset by aggregating available landmark lines on the track. Subsequently, the author generated a training dataset by introducing randomly generated missing data points and trained it using a random forest model, resulting in a notably effective prediction model.

During the exploration phase, the author investigated three key aspects: the quantity, positions, and distribution of known data within the input sequence. It was deduced that input sequences conducive to producing accurate prediction results possess the following characteristics: a substantial quantity of known data, a dispersed distribution of known data, proximity of known data positions to each 10-meter interval, and a preference for known data within the initial 50 meters over the latter 50 meters. These guidelines empower users of the model to optimize its utility and obtain satisfactory prediction outcomes.

This research addresses a significant gap in the field of predicting segmented data for 100-meter races by enabling the acquisition of complete segmented data from any known data position, thereby alleviating the prevailing challenge of data scarcity. However, as outlined in the Discussion section, potential overfitting issues necessitate

further refinement and exploration by future researchers employing larger and more diverse datasets.

References

1. Hua, Z.. Research on Su Bingtian's Technical Characteristics in the 32nd Tokyo Olympics. Zhongbei University, 2023.
2. Majumdar, A. S., & Robergs, R. A. The Science of Speed: Determinants of Performance in the 100 m Sprint. International Journal of Sports Science & Coaching, 2011 6(3), 479-493.
3. Pang, L.. Kinematic Analysis of Key Technical Aspects of the Top Eight Athletes in the Men's 100m Final at the 2019 Doha World Championships. Capital University of Physical Education and Sports, 2021.
4. Graubner, R., Buckwitz, R., Landmann, M., & Starke, A. Scientific Research Project: Biomechanical Analyses At the 12th IAAF World Championships in Athletics 2009. Final Report. Sprint Men. Darmstadt: DLV, 2009.
5. Wang, H.. Comparative Study of Su Bingtian and Bolt's 100m Running Technique. Chengdu Sports University, 2017.
6. Xie, Z.. Plan of Standard Outdoor Track Markings for 400 m by the International Association of Athletics Federations (IAAF)Athletics, 2017(10):32-33.
7. Louppe, G., "Understanding random forests: From theory to practice." arXiv preprint arXiv:1407.7502, 2014.
8. Sun, Z., Wang, Guotao, Li, Pengfei, Wang, Hui, Zhang, Min, Liang, Xiaowen, "An improved random forest based on the classification accuracy and correlation measurement of decision trees." Expert Systems with Applications, Volume 237, Part B, 2024, 121549.
9. Marmolin, H. "Subjective MSE Measures," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 16, no. 3, pp. 486-489, May 1986..
10. Power, A., Burda, Y., Edwards, H., et al. "Grokking: Generalization beyond overfitting on small algorithmic datasets." arXiv preprint arxiv:2201.02177, 2022.