John Smutny
ECE5424 - Advanced Machine Learning
Dr Yue Wang
10/05/2022

Traits of the Best Athletes:

Men’s Decathlon and Women’s Heptathlon

1 Introduction

There are many ways to judge an athlete in sports; how high they jump, their speed, strength, or solely by their results. News talk radio and television have endless debates about which athlete is the best in their profession, who is better than whom, who is the greatest of all time. In Track & Field, the Men’s decathlon and Women’s heptathlon competitions decide who is the best overall athlete over two days of numerous events that test their speed, strength and skill. Starting with Jim Thrope in 1912, the Olympic decathlon winner began being referred to as “the greatest athlete in the world” [1]. Please see the section “*Background: Format of the Decathlon/Heptathlon*'' for more information on the technical details of the competition.

This paper aims to find patterns and relationships in recorded individual performances (and other factors) that can help determine what makes the best decathlon athlete. Each individual’s performance at the Olympic games from 1912 to 2020 will be used to train an unsupervised learning model that will organize these athlete’s into common clusters. These clusters will be used to better understand the decathlon/heptathlon.

Definitions

- Contest: a track & field tournament that hosts events other than the decathlon/heptathlon.
- Competition: Description of all decathlon/heptathlon events in one contest/tournament.
- Event: one specific discipline of the decathlon (10) or heptathlon (7).

2 Aims and Curiosities

This paper aims to use classifications from unsupervised ML models to help gain insight into the the following questions about decathlon/heptathlon athletes and competitions:

1. Understand if there are a variety of athletes in the decathlon or one homogeneous framework.
2. Do athletes fall into categories similar to the events: strength vs speed vs explosiveness?
3. (if possible) How do biometric traits (height/weight) define certain subsets of competitors?
4. Can similarly clustered countries tell us about those regions/countries?
5. Do the patterns in the decathlon also appear in the heptathlon event?
6. Other items of interest:
    a. Any influence of home-field advantage/pride. Do athletes from the host country cluster in any way over Olympic history?
    b. By analyzing historic performances by each event in isolation: Do certain countries consistently perform better at the decathlon/heptathlon competition or a specific event?

3 Prior Experiments and Analysis

No Machine Learning applications around the Olympics were found at the writing of this proposal. However, there was an analysis of the Decathlon event specifically that can provide context to the eventual ML model's output. In 2012 (revised in 2021), Professor John Barrow of the University of Cambridge's Mathematics department did a mathematical analysis of how the decathlon's scoring system incentives certain types of athletes and certain events over others. This is based on the constants used in the points equation for each event. Each equation has a 'Power Index' multiplier that magnifies an athlete's improvement. IE, a larger multiplier means that a subtle improvement in that event leads to more points then an event with a smaller multiplier. The three events with the largest multipliers (highest point potential) were the Long Jump, 110m Hurdles, and 100m Sprint. While the events that had the lowest multiplier (lowest point potential) were Shot Put, Discus, and Javelin and 1500m Run. Therefore, according to Prof Barrow's analysis, the scoring incentives would lead the best decathlon athletes to prioritize sprints, jumps and then throwing events (in that order) to maximize point potential. [2]

4 Datasets

This research will rely upon two primary datasets of Olympic competitors:

1. The first dataset is web scraped data from Olympedia.org's database detailing every athlete's performance in each Olympic games from 1912 to 2020 [3]. These performance metrics include their final score of the competition and each event's result. Therefore, the point value scored in each event can be calculated or the raw measure can be used out right. This dataset is collected by using the BeautifulSoup web-scraping public python package. The list below shows the resulting features from the web-scraping. In private proof-of-concept experiments, attempts to extract the Olympedia.org HTML code into a csv available for ML processing have been successful. The features of this dataset are shown below.

   { Position, Bib Number, Competitor, Country, Points, 100 Meters, Long Jump, Shot Put, High Jump, 400 Meters, 110 Meters Hurdles, Discus Throw, Pole Vault, Javelin Throw, 1500 Meters, Medel [Gold, Silver, Bronze] }

2. The second dataset is from a contributor to Olympedia.org (Kaggle user rgriffin described in the last section), that is an archive of each athlete's recorded height and weight during each Olympic games [4]. This dataset is provided as a ready to use csv file. The features of this dataset are shown below.

   { ID, Name, Sex, Age, Height, Weight, Team, Country, Games, Year, Season, City, Sport, Event, Medal }

   There are two challenges with this dataset in particular. First, provided data entries will have to be matched to athlete data from the first dataset based on A) the athlete's name and B) the year of the recorded competition. Second, the author's statistical analysis noted that athlete height and weight measurements were not significantly recorded (over 50% participation) until the 1960 Olympic games [4]. Therefore, caution must be applied when using these biometric markers and any analysis including height and weight will be restricted from 1960 to modern day.

5 ML Model Design

In order to learn more about the decathlon athletes, the datasets described in the previous section will be processed by an unsupervised ML model to organize and relate athletes together based on their event performances. Additional factors could include; athlete height, weight, and if they are from the host nation. Some example algorithms that can be considered for unsupervised clustering include Hierarchical Divisive Clustering, kMeans learning and multi-class Support Vector Machines classifiers.

5.1 Additional Methods

The dimensionality reducing method, Principal Component Analysis (PCA), is planned to be used later in the analysis to simplify the clustering process. Intuitively, the decathlon events could be reduced into their specific disciplines (running, jumping, throws) to see if this reduction aids in athlete clustering. Once finished, external validation, cluster cohesion, and cluster separation metrics will be the primary method of validating each cluster and extracting that cluster's meaning [6]. The methods and ML model could change based on metric feedback.

6 Potential Challenges

Some challenges that must be overcome in this project focus on data modifications and considering the human nature of the decathlon/heptathlon competition format.

1. Re-formatting the scraped data of Olympic results into a suitable format for the ML model. Also includes adding features as desired.

2. Aligning the results and biometrics datasets to match each athlete for each competition. See the "*Datasets*" section for more details. If the 'biometric' dataset is unusable, that significantly hinders this analysis.

3. Deciding what scoring table to use when comparing athlete's from different Olympic games or if they should be kept as raw distance/time measurements. Professor John Barrow's report showed how different scoring systems allow for different results. The performance dataset from Olympedia.org does provide Point totals for various scoring tables; such as 1934, 1952, 1962, 1985, and modern day. In addition, would normalizing each score based on the Olympic games improve clustering. IE: no matter the scoring table, that decathlon's Gold medalist's total points act as the linear normalized maxima. That way athletes from all scoring tables are on the same playing field and only their measures are considered.

4. Properly understanding the resulting clusters. After modeling the dataset in different ways, can ML help us understand the relationships inside the decathlon/heptathlon events.

References

[1] Flatter, R. (n.d.). Thorpe preceded Deion, Bo (ESPN, Ed.). *ESPN Sports Century*, (No7). https://www.espn.com/sportscentury/features/00016499.html

[2] Mallon, B. (2006). *Olympedia* (International Society of Olympic Historians, Ed.) [Database of complete Olympic results and Olympians]. OlyMADMen. https://www.olympedia.org/

[3] Barrow, J. (2012). Decathlon: the Art of Scoring Points (University of Cambridge Faculty of Mathematics, Ed.). *NRICH*. https://nrich.maths.org/8346

[4] rgriffin. (2018, July 27). *Olympic history data: thorough analysis* (27). Kaggle.com. https://www.kaggle.com/code/heesoo37/olympic-history-data-a-thorough-analysis/report

[5] IBM. (2020, September 21). *What is Unsupervised Learning?* IBM. Retrieved September 27, 2022, from https://www.ibm.com/cloud/learn/unsupervised-learning

[6] Remy, M. (2020, January 27). *Unsupervised Machine Learning: Validation Techniques*. Guavus. Retrieved September 27, 2022, from https://www.guavus.com/technical-blog/unsupervised-machine-learning-validation-techniques/

Appendix

A. Timeline

Week #	Date	DATA	MODELING
1	10/17/2022	1. Create a 'synthetic' dataset that is in the predicted end format of web scraped data. 2. Web scrape 'results' dataset into a usable format. 3. Perform statistical analysis of the dataset to make a plan to clean data if necessary. 4. Web scrape 1st set of 'biometric' dataset.	
2	10/24/2022	1. Attempt to map the 'biometric' subdataset to 'results' dataset. If successful… 2. Web scrape the rest of the 'biometric' dataset 3. Perform statistical analysis of the dataset to make a plan to clean data if necessary.	Begin **kMeans Cluster** unsupervised ML model using the 'synthetic' dataset
3	10/31/2022	Attempt to map the full 'biometric' dataset to 'results' dataset.	Begin **Divisive Hierarchical** unsupervised ML model using the 'synthetic' dataset
4	11/7/2022	Make a judgment call on if 'biometric' data is usable.	Begin **Support Vector Machine** unsupervised ML model using the 'synthetic' dataset Analyze results.
5	11/14/2022		1. Model 'full' dataset with all previous models. Analyze results.
6	11/21/2022		2. Apply **Principal Component Analysis** to an unsupervised ML model. Analyze results 2. Repeat procedure for heptathlon dataset. Analyze results.
7	11/28/2022		Final Preparations
8	12/5/2022	REPORT	

9	12/12/2022	REPORT	

B. Background: Format of the Decathlon/Heptathlon

In the Summer Olympics, the decathlon (men) and heptathlon (women) are unique contests in the ‘Athletics’ category that involve multiple events over a two day period. The same athletes compete in all events and are given a score based on their individual distance/time. The score for each event is decided by a mathematical “scoring table” and not in comparison to how other competitors perform (1st, 2nd, 3rd, etc). This “scoring table” has changed at various points in Olympic history based on different criteria, and not all events are given the same weight [3]. The men’s decathlon involves 10 events (100M, Long Jump, Shot Put, High Jump, 400M, 110M Hurdles, Discus, Pole Vault, Javelin, 1500M) while the women’s heptathlon involves 7 events (100M Hurdles, 200M, 800M, High Jump, Long Jump, Shot Put, Javelin).

Decathlon		Heptathlon	
Day 1	Day 2	Day 1	Day 2
100 m	110 m Hurdles	100m Hurdles	Long jump
Long jump	Discus	High jump	Javelin
Shot put	Pole vault	Shot put	800m
High jump	Javelin	200m	
400 m	1500m		