Traits of the best athletes: Men’s Decathlon and Women’s Heptathlon

John Smutny

ECE5424 - Advanced Machine Learning

Dr Yue Wang

09/27/2022

**Abstract**

There are many ways to judge an athlete in sports; how high they jump, speed, strength, performance in their sport. News talk radio and television have endless debates about which athlete is the best in their profession, whether that be scoring goals, dunking a basketball, or running around a defender. In Track & Field, the Men’s decathlon and Women’s heptathlon competition decides who is the best overall athlete over two days of numerous events testing speed, strength and skill. The winner of the 1912 Olympic Decathlon, Jim Thrope, was called “the greatest athlete in the world” by the Swedish king at the time [1]. This paper aims to find patterns and relationships in recorded individual performances that can help determine what makes the best 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 in common clusters. Those clusters may or may not include the athlete’s height and weight.

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).

**Background**

The Olympics is a sports competition that started in Athens, Greece and involves many different events and competitors. The modern Olympics refer to either the Summer or Winter games that both occur every four years in a single nation around the world. 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 performance. The score for each event is decided by a mathematical “scoring table” and not in comparison to other competitors [2]. This “scoring table” has changed at various points in Olympic history based on different criteria, but 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).

**Specific Aims and Goals**

This paper uses an unsupervised Machine Learning model and other techniques to find relationships between decathlon and heptathlon athlete performance and their overall performance. These relationships are planned to be found by looking at the athlete’s individual performances in each event and other factors with how they performed overall. Some specific questions include

1. Determine if there are any strategies to win the decathlon based on individual event performance. (specialize in power vs speed vs explosive events, be best at the events latest in the day, be younger/older, etc). Should athletes attempt to excel at more events or just the highest scoring ones?
2. What are the traits of a successful athlete: strength vs speed vs explosiveness.
3. Do any groups of countries fit into specific clusters more often than others?
4. Do athletes from the host country perform better than other athletes?
5. Do certain countries consistently perform better at the decathlon/heptathlon competition or a specific event?
6. (if possible) Which biometric traits (height/weight) result in the best success per event and overall?

**Prior Experiments and Analysis**

No Machine Learning applications around the Olympics could be found at the writing of this proposal. There were two separate statistical analyses of the Olympic games and of the Decathlon event in general that can provide data and context to any ML model/analysis.

1. In 2012 (then revised in 2021), Professor John Barrow of the University of Cambridge’s Mathematics department did a mathematical analysis at how the decathlon’s scoring system incentives certain types of athletes and prioritizes some events over others. They showed how the current 2001 score system prioritizes certain events over others and which events provide the most points compared to others. The top three events with the highest ability to score points were the Long Jump, 110m Hurdles, and 100m Sprint. While the lowest events to score points are the Shot Put, Discus, and Javelin and 1500m Run. Therefore, based on incentives, the best scoring decathlon athletes would have the best running and jumping scores. [2]
2. In 2018, Kaggle user (and now contributor to Olypedia.org) rgriffin performed a general statistical analysis of the athlete’s and countries throughout Olympic history. In this analysis, rgriffin provides one of the two datasets used in this research. Their analysis describes and diagrams the history of the Olympics, including geographic representation, medal counts, number of athletes, and how athlete’s body’s have changed over time. Specific to this paper is weight and height. Specifically, rgriffin’s statistical analysis of the Olympic biometric data, they describe that the height & weights of each athlete did not reach a significant number until the 1960 Olympic games (see figure below from rgriffin). Therefore, caution must be taken when applying biometric markers to any ML model. [4]

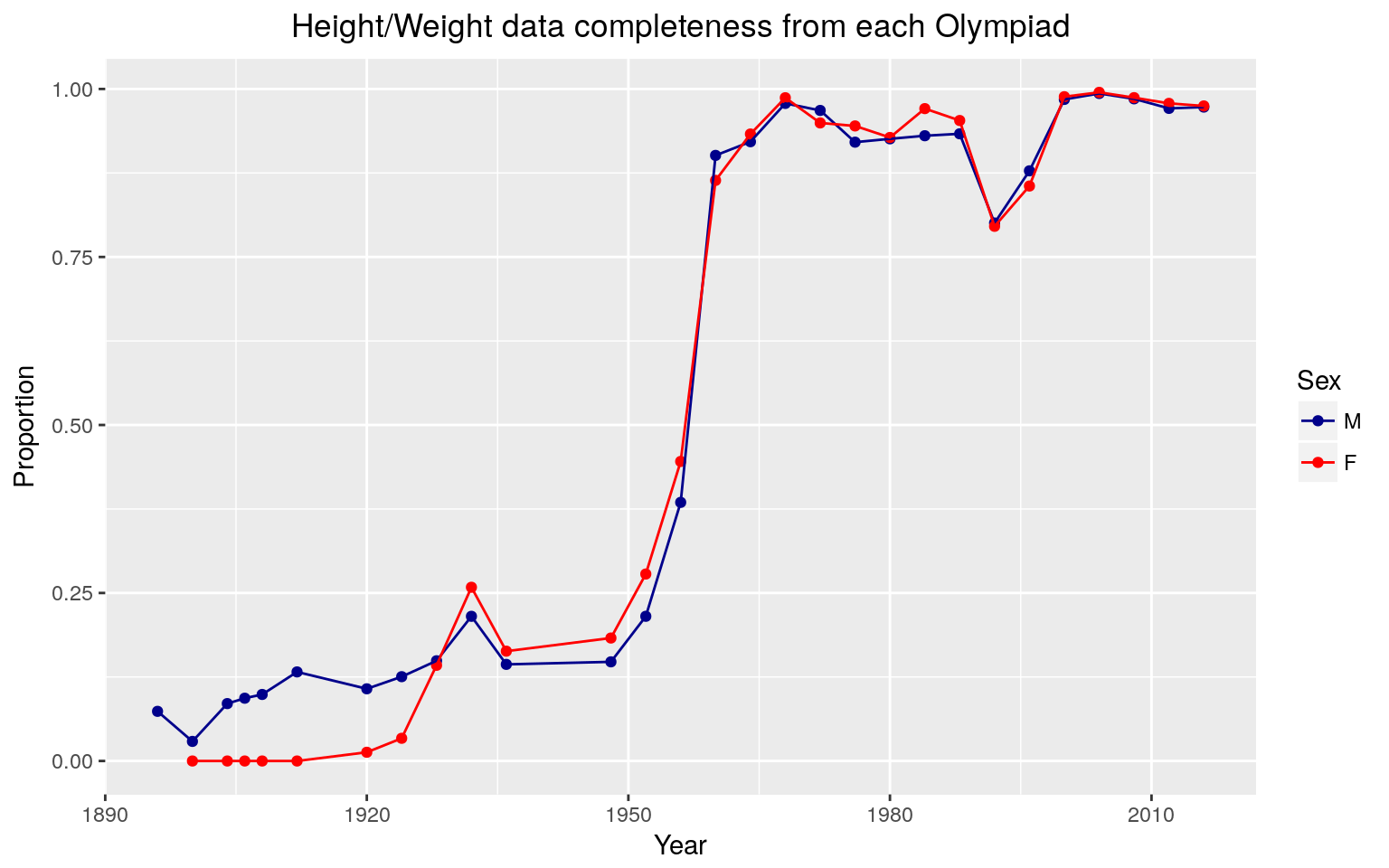


Figure D&E1: Data Completeness chart from Kaggle user rgriffin’s statistical analysis [4]

**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 the final score of the competition and each event’s results. Therefore, the point value scored in each event can be calculated. This dataset was collected by using the BeautifulSoup web-scraping public python package. The list below shows the resulting datasets from each web-scraping. In private code experiments, attempts to extract the HTML code into a csv available for ML processing has been successful.

{ 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] }

1. The second dataset is from a contributor to Olympedia.org (Kaggle user rgriffin), but is an archive of each athlete’s recorded height and weight for that particular Olympic games [4]. It has been noted however that the weights and heights of athletes were not significantly recorded until the 1960 games. Therefore, caution must be taken when applying biometric markers to any ML model. This dataset is provided as a ready to use csv file.

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

**Design and Methods**

In order to learn more about decathlon athletes, the datasets described in the next section will be processed by an unsupervised ML model to organize and relate athletes together based on their event performance and (if time) their height & weight. Specifically, the divisive clustering model is planned to be used, because it provides a visual separation of each cluster from the original dataset (see figure below) [5]. Each step and separation of the data will provide more information as to why certain clusters came together compared to others. This is thought to help answer the paper’s questions regarding athletic performance and external relationships. Another KMeans model could be made to answer more specific questions about a preference to one type of events (running, jumping, throws).

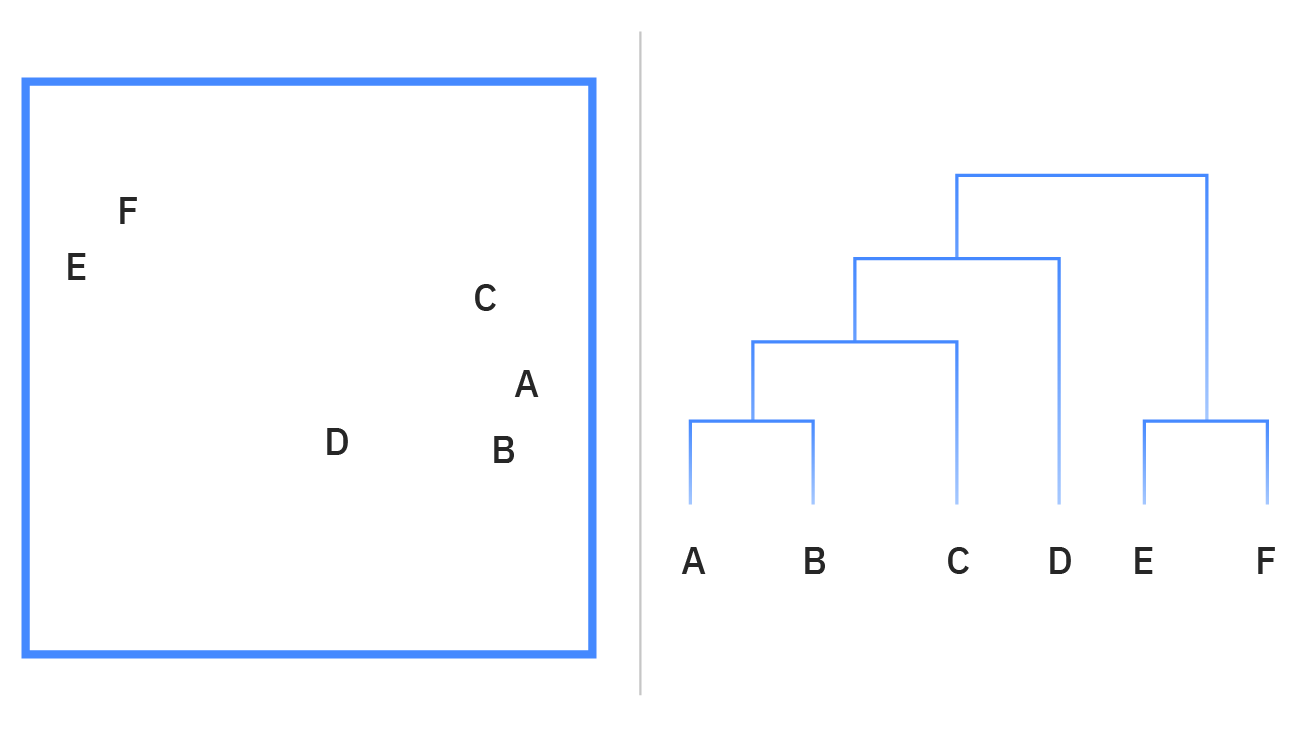


Figure D&M1: Outline of Divisive clustering. (left) 2D Clustering. (right) Resulting hierarchical path to each cluster. [5]

Based on model performance, ‘soft’ KMeans Clustering and Gaussian Mixture models may be considered in lieu of poor results from the divisive clustering model. Additionally, the dimensionality reducing method, Principal Component Analysis (PCA), is planned to be used later in the process to improve clustering. Intuitively, the decathlon events could be reduced into their specific disciplines (running, jumping, throws) to see if it aids in athlete clustering. Once finished, external validation will be the primary method of validating each cluster and extracting that cluster’s meaning. This observable validation of each cluster will be guided by cluster cohesion and cluster separation metrics [6]. The methods and ML model will change based on metric feedback.

**Planned Challenges**

Some major challenges that must be overcome in this project include data modifications and the human nature of the decathlon/heptathlon competitions.

1. Re-formatting the data from Olympic results dataset so that it is in a usable way after it has been scrapped from the internet. These modifications are outlined in the next section.
2. Aligning the results and biometrics datasets to match each athlete for each competition. IE. Ensuring that athlete bio information can match athlete competition results for the correct competition. The original performance dataset scrapped from Olympedia.org provides information on all Olympic games and in theory aligns with the biometrics dataset gathered by Kaggle user rgriffin. Each athlete in the biometrics dataset will have to be matched with the performance dataset to make the athlete height and weight usable. If there is a significant issue matching the athlete’s height and weight information with their performance, then this paper will only consider an athlete’s performance in the ML model.
3. Deciding what scoring table to use when comparing athlete’s from different Olympic games. 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.
4. Based on what scoring table is used, decide if each event’s metrics are left in terms of the raw distance/time or if it should be converted into a point value corresponding with what scoring table is used by the ML model.

**Planned Data Modifications**

To provide the best ML model and analysis, several steps will have to be done to prepare the data for the unsupervised learning model.

1. The decathlon web scraped results will have to be formatted in a usable format.
   1. All competition Points are in integer formats.
   2. Convert the “Medal” feature into a ordinal integer { 0, 1, 2, 3 }
   3. Create a “host country” flag to indicate if the athlete is from the same country that is hosting that Olympic games.
   4. (pending) Convert all event records in point values
2. Combine datasets
   1. Compare the athlete’s name and competition year between datasets. If there is a match, join the datasets.

**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/