Classifying Olympic Athletes By Sport and Event

Matt Dickenson
Department of Political Science
Duke University
Durham, NC 27708
mcd31@duke.edu

Abstract

Can the sport or event an Olympic athlete participated in be predicted from her height, weight, age, and sex? One sports analyst has claimed that he could do this with a high degree of accuracy, as part of a broader argument that genetics determine sporting success. In contrast, another school of thought argues that an athlete's training (on the order of 10,000 hours) is more predictive. This project classifies athletes from the 2012 Olympic games using hierarchical clustering, conditional inference trees, evolutionary trees, random forests, and neural networks. Between 25 and 30 percent of observations are classified accurately. This suggests that physical features play a strong predictive role, but training likely explains a large proportion of the remaining variance. Above a minimum threshold of physicality, training plays an important role in sporting success.

1 Introduction

1.1 Motivation

To what extent do biological traits determine sporting success? At the highest level of amateur sports–the Olympic games–we see differences in the physical characteristics of participants across sports. Can these differences be exploited to classify individuals by sport or event given their physical attributes?

This project was inspired by a claim made by David Epstein, author of *The Sports Gene* [**?**]. This claim is expressed in an interview with Russ Roberts:

> **Roberts**: [You argue that] if you simply had the height and weight of an Olympic roster, you could do a pretty good job of guessing what their events are. Is that correct?
> **Epstein**: That's definitely correct. I don't think you would get every person accurately, but... *I think you would get the vast majority of them correctly.* And frankly, you could definitely do it easily if you had them charted on a height-and-weight graph, and I think you could do it for most positions in something like football as well.[1]

Epstein's work is in large part a counter-argument to the "10,000 hour rule," popularized by Malcolm Gladwell, which claims that that amount of practice is necessary to attain mastery of a skill [**?**]. This project analyzes whether an athlete's sport or event can be accurately predicted by their physical features. If true, that suggests that athletes choose the event that best leverages their natural predisposition. If not, that would suggest that athletic ability can be understood as a latent trait that can be applied to the sport of one's choice. The remainder of this paper describes related work, outlines the machine learning methods used, and presents the results.

1.2 Related Work

2 Model and Methods

2.1 Problem Definition and Data Sources

The goal of this project is to predict an athlete's Olympic event given their height, weight, age, and sex. Given data $\mathcal{D}$ on these $p = 4$ features our goal is to predict athletes' sports or events $\mathbf{y}$ (which are nested in sports).[2] The quantity we are interested in, then, is $p(\mathbf{y}|\mathcal{D})$.

Data was obtained for 10,383 participants in the 2012 London Olympics from *The Guardian*.[3] The processed data consists of 8,856 complete cases. Of these, 6,956 participants were split into training (n=3,520) and test (n=3,436) sets for classification by sports. In 2012 there were 27 sports in the Summer Olympic Games, excluding participants in the "Athletics" sports category. Because Athletics is such a large category (with 1,900 participants and 48 events) with participants' exhibiting a wide range of body types, it tended to dominate the classification models when it was included. Omitting Athletics participants from the sport classification task substantially improved accuracy. The remaining 1,900 Athletics participants were split into training (n=907) and test (n=993) sets for classification by event.

2.2 Features of the Data

For both classification tasks participants' height, weight, age, and sex were used as features. Some sports and events exhibit relatively well-clustered features, whereas others are less clearly defined. In the sport classification task, participants in archery, handball, swimming, and triathlon exhibited similar features–and were thus difficult to classify accurately–while basketball players, rowers, weightlifters, and wrestlers had more distinct features (see Figure 1. For event classification it was difficult to classify athletes in the 100m race, 400m hur-

[1] An audio recording of the interview, as well as a partial transcript, is available at `http://www.econtalk.org/archives/2013/09/david_epstein_o.html`.

[2] In the Olympic games, "sports" are somewhat broad categories, such as athletics and weightlifting, whereas "events" refers to specific competitions such as the men's 100-meter race.

[3] `http://www.theguardian.com/sport/series/london-2012-olympics-data`

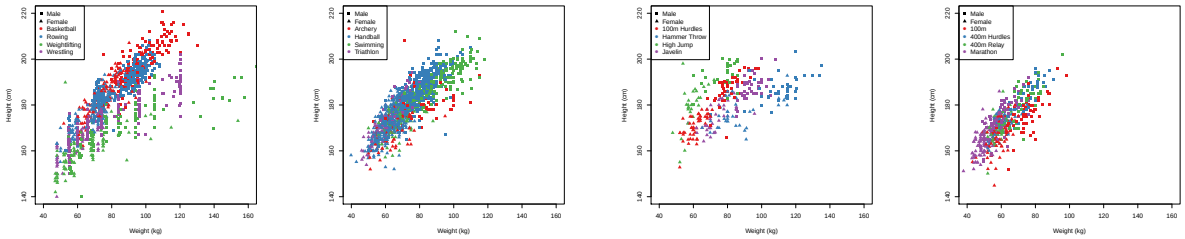

Figure 1: Comparing well-differentiated and poorly-differentiated feature clusters for sports (left two panels) and events (right two panels).

dles, 400m relay, and marathon, but easier to classify participants in the 100m hurdles, hammer throw, high jump, and javelin events.

2.3 Model

In order to classify an athlete from the test set, we can use the posterior predictive distribution

$$p(\mathbf{y}|\mathcal{D}) \quad = \quad \sum_{k \in N} w_k p(\mathbf{y}|\mathcal{D}_k),$$

where w_k is the weight on category k, $\mathcal{D}_k$ is the feature data for observations in category k, and N is the set of categories. Categories can represent sports or events depending on the classification task. The substantive interpretation of the posterior predictive distribution is the probability that a given athlete belongs to a certain category.

2.4 Machine Learning Methods

Several machine learning methods were used to classify Olympic athletes by sport and event. Hierarchical clustering was performed with a Gaussian likelihood model, with the covariance matrix $\Sigma_k = \lambda_k D_k A D_k^T$ (an ellipsoidal distribution with variable volume, equal shape, and variable orientation) [**?** **?**]. Conditional inference trees were formed by recursive binary partitioning, using the `party` package in $\mathcal{R}$ [**?** **?**]. Evolutionary trees were constructed to be globally optimal by minimizing the misclassification rate, using the `evtree` package [**?**]. Breiman's Random Forest algorithm was used with 500 trees, as implemented in the `randomForest` package [**?**]. Single-hidden-layer neural networks were constructed with 30 units in the hidden layer for sport classification and 50 for event classification, with the package `nnet` [**?**].

3 Results

3.1 Classification by Sport

Figure 2 presents classification matrices for sports (left two columns) and events (right two columns). Each matrix is row-normalized, so the darkest cell on a row indicates the modal predicted class (columns) for true members of the class (rows). Dark cells along the diagonal indicate accurate predictions. A legend is provided in Figure 3. Hierarchical clustering was not used for the event classification task (see Section 3.2).

3.2 Model Diagnostics

Table 1 presents the accuracy of each model for the sport and event classification tasks. Hierarchical clustering could not be used for event classification due to the small sizes of the clusters (for example, there was complete data for only one participant in the women's triathlon event). The ratio columns indicate how well the model performs on the test data relative to the training data. Values near one mean that the model does about as well out-of-sample as it does on the training data, while values near zero suggests a model that is overfit to the trianing data.

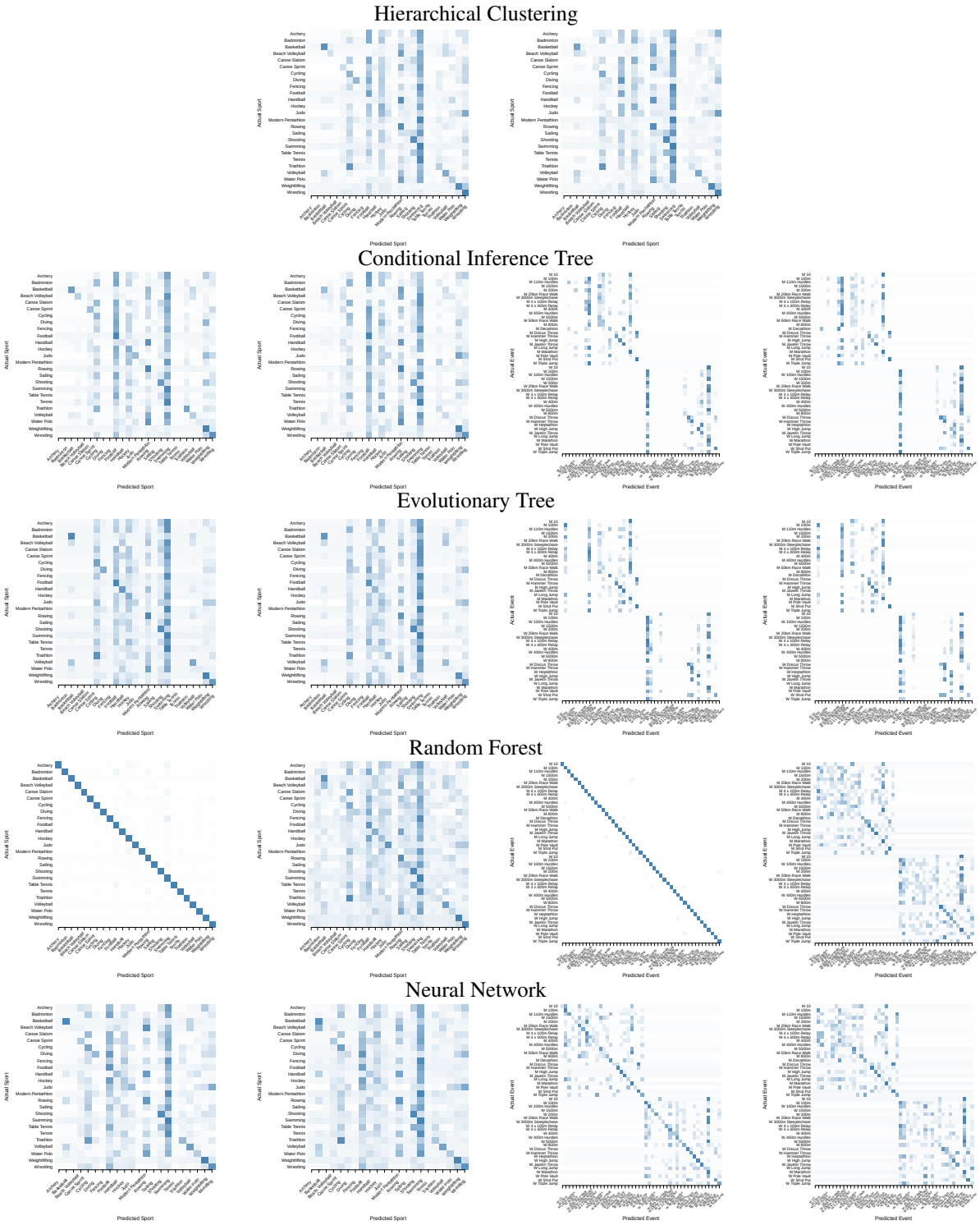

Figure 2: Classification matrices by method for sports (left two columns) and events (right two columns). Participants' actual sports/events are labeled on the rows of each image, and the predicted sports/events are labeled on the columns. The cells represent row-normalized frequencies. Darker shades on the diagonal indicate accurate classification.

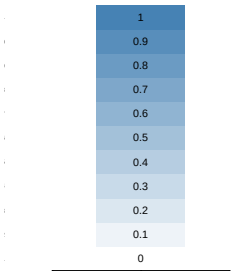

Figure 3: Legend for matrices in Figure 2.

Table 1: Accuracy for training and test sets.

	Sports			Events		
	Train	Test	Ratio	Train	Test	Ratio
Hierarchical Clustering	.272	.271	.998			
Conditional Inference Tree	.279	.219	.784	.277	.218	.787
Evolutionary Tree	.292	.236	.807	.303	.230	.757
Random Forest	.923	.244	.265	.976	.228	.233
Neural Network	.280	.265	.949	.397	.249	.623

3.3 Interpretation

Hierarchical clustering had the best test set performance, suggesting a robust model that predicts moderately accurately without overfitting the training data. Evolutionary trees and neural networks both exhibit higher accuracy for events than sports, and maintain acceptable out-of-sample accuracy. Conditional inference trees do slightly less well for both training and test sets. Random forests tend to overfit the training data but still do well on the test set. Overall, this appears to be a difficult classification problem.

4 Conclusion

The results presented above show that classifying athletes by sport and event can be achieved with moderate accuracy using only a few features. However, Epstein's confidence that he could accurately classify Olympians by only height and weight seems misplaced. Additional features such as arm length and torso length could improve predictive accuracy. Traits of athletes in some sports and events exhibit noticeable clustering, while other categories are less distinct (i.e. multi-modal). Athletes in some sports and events have a well-defined body type, but Olympians exhibit a wide range of physical features. As the Olympics become more competitive, training will continue to play a vital role in sporting success.

Acknowledgments

Thanks to Michael D. Ward and National Science Foundation Grant #3331808 for support during this project. Any conclusions or errors are the sole responsibility of the author.

References