

陸上トレーニングデータと大会結果に関するデータ分析

2020SS014 堀江爽太郎

指導教員：小市俊悟

1 はじめに

陸上競技の一つに 1000m 走という種目がある。しかし、オリンピックや世界陸上に 1000m という種目はなく、近い距離の種目としては 800m 走と 1500m 走という種目しかない。800m 走には短距離型の選手が多く、1500m 走には長距離型の選手が多いため、1000m 走は、それらの選手が混在する境界的な種目となっており、選手が短距離型か長距離型に合わせて、適切に練習しなければならない種目でもあると考える。練習で得られるデータを対象に、4 種の回帰分析を行う。これらの分析を通じて、どのような項目が 1000m 走のタイムに影響を与えているかを分析する。続いて、1000m 走では、選手の特徴に合わせた練習が重要であるので、練習データから選手の特徴を把握することを目的に、クラスター分析を行う。

2 データ

表 1 に示すデータを収集した。

表 1 データ

データの項目	詳細
time	練習中に計測した 1000m 走のタイム (s)
sb1	800m の過去 1 年間のベスト記録 (s)
sb2	1500m の過去 1 年間のベスト記録 (s)
sb3	5000m の過去 1 年間のベスト記録 (s)
fatigue	足の疲労度 1 が低く 5 が高い
condition	体調 1 が悪く 5 が良い
distance	1 週間の走行距離 (km)
updist	ウォームアップの距離 (km)
uptime	ウォームアップのペース (s)
bpm	ウォームアップ時の心拍数 (回/分)
tem	練習時の気温 (°C)

これらのデータのうち、sb1, sb2, sb3 については練習までに実施された大会記録等に基づく。distance は週間走行データを見て記入する。その他のデータは、練習ごとに記録される。練習中は、心拍数等計測機能付き GPS ランニングウォッチを付けてもらい、その記録から updist, uptime, bpm を記入する。fatigue と condition については対象者本人の主観的な感覚に基づき、記載してもらう。練習で 1000m を走り、time と tem を記入する。表 2 は、これらのデータの一例である。

表 2 練習データの例

データの項目	time	sb1	sb2	sb3
選手 1	166.8	121.62	246.6	975
データの項目	fatigue	condition	distance	updist
選手 1	4	3	65.15	3.39
データの項目	uptime	bpm	tem	
選手 1	258	150	22	

3 回帰分析

3.1 回帰分析の方法

どの項目 (変数) が 1000m 走のタイムに影響を与えるのかを明らかにするために、被説明変数を time とし、説明変数を他の変数とする回帰を行った。回帰の方法は、線形回帰、非線形回帰、決定木回帰、ニューラルネットワーク (以後 NN と略す場合がある) による回帰の 4 種類である。非線形回帰は、2 次と 3 次の多項式による回帰とする [1]。決定木回帰では、木構造の各頂点に説明変数 (のいずれか) に対する条件が定められた回帰木を作成することになる。各データには、回帰木の根に定められた条件がまず適用されるが、その真偽によって、次に適用すべき条件が木構造に従って定まる。このような条件の適用を木構造の葉に達するまで行い、最終的に葉に割り当てられた値が回帰木による予測となる。ニューラルネットワークによる回帰には中間層が二層のニューラルネットワークを用い、それらの層のニューロン数はそれぞれ 512 と 511 とした。この値は試行錯誤の中で最も予測精度が高かったものである。これらの分析は scikit-learn が提供している下記のモジュールを用いたものである。

- 線形回帰:LinearRegression
- 非線形回帰:PolynomialFeatures
- 決定木回帰:DecisionTreeRegressor
- ニューラルネットワーク:MLPClassifier

4 実験結果

線形回帰、非線形回帰、決定木回帰、ニューラルネットワークを適用した結果について、選手 1 を例に表 3 に示す。

表3 各回帰の結果

回帰の種類	実際のタイム	予測タイム	誤差
線形回帰	166.8	168.2	1.4
2次非線形回帰	166.8	166.8	0
3次非線形回帰	166.8	166.8	0
決定木回帰	166.8	165.5	1.3
NN	166.8	169.0	2.2

表4 分析手法ごとの標準偏差

データの項目	標準偏差
線形回帰	3.35
2次非線形回帰	1.31
3次非線形回帰	0
決定木回帰	2.18
NN	3.31

4.1 回帰分析の考察

線形回帰では次の回帰式が得られた。
 $(\text{time})=0.13(\text{sb1})+0.72(\text{sb2})+0.054(\text{sb3})+1.62(\text{fatigue})$
 $+0.40(\text{condition})-0.023(\text{distance})+0.61(\text{updist})$
 $+0.031(\text{uptime})+0.0071(\text{bpm})-0.13(\text{tem})$
この式から800mのタイムと足の疲労度の項目が大きく影響を与えていることが分かった。非線形回帰に関しては2次と3次で行った。3次に関しては誤差0となるが、明らかにデータに合わせすぎた結果になってしまった。決定木回帰は標準偏差では2.18となった。決定木回帰が精度が高い要因として選手を長距離型と中距離型のタイプ別に分けるような回帰木を作成できたからであると感じた。ニューラルネットワークは適切なニューロン数を見つけるのが難しく、多く試した結果、中間層が二層で512, 511のものが最も標準偏差が小さくなった。

5 クラスタ分析

5.1 クラスタ分析の方法

次に、選手を長距離型や短距離型などという分類よりさらに詳細に分類することを目的にクラスタ分析を行った。クラスタ分析には非階層的クラスタリング k-means 法を本研究では用いる。

5.2 クラスタ分析の結果

表5は、選手を5つのクラスタに分けたとき、それぞれのクラスタにおける各項目の平均値を示したものである。

5.3 考察

1番目のクラスタは、800mの記録が一番速く、1500mのタイムも1番目に速いことからスピードタイプのランナーの集まりと言える。走行距離も短く、心拍数も低いことから中距離選手の集まりだと言える。2番目のクラスタは、1000mのタイム、800m、1500m、5000mのSBが

表5 クラスタ分析の結果

データの項目	time	sb1	sb2	sb3
クラスタ1	167.5	245.4	121.3	949.8
クラスタ2	173.6	262	127.9	1005
クラスタ3	164.6	247.8	124.9	909.8
クラスタ4	167.2	248.4	123.5	949.1
クラスタ5	196.5	291.1	139.9	1108.5
データの項目	fatigue	condition	distance	updist
クラスタ1	3.0	2.5	41.7	2.0
クラスタ2	2.8	3.65	53.3	2.56
クラスタ3	2.6	3.7	125.6	3.49
クラスタ4	2.9	3.5	43.4	2.54
クラスタ5	3.2	3.46	22.0	2.05
データの項目	uptime	bpm	tem	人数
クラスタ1	351	137.2	22	6
クラスタ2	274.2	144.2	28	20
クラスタ3	252.2	139.0	26.6	10
クラスタ4	259.0	143.1	24.0	34
クラスタ5	311.6	147.3	28.1	11

4番目であるが、走行距離は2番目であることから能力は高くないが練習量は比較的多くしている選手の集まりと言える。3番目のクラスタは、1000m、5000mのタイムが最も速く、走行距離も秀でて多いことから一番能力が高くストイックな選手の集まりであるといえ、実業団選手に近い数値でもある。4番目のクラスタは、1000mのタイム、800m、1500mのSBが2番目であり、幅広い距離の種目を走れるオールラウンダーのランナーの集まりと言える。5番目のクラスタは、1000mのタイム、800m、1500m、5000mのSBが5番目であり、走行距離も少なく、アップの心拍数も上がりやすいことから能力、努力ともに劣った選手の集まりと言える。

6 おわりに

分析を始めた当初の目標である2.5秒以内の精度を非線形回帰と決定木回帰で達成することができた。ニューラルネットワークの精度は3.31秒と目標には届かなかったが、使用当初よりは2秒ほど精度を高めることができた。クラスタ分析を行うことで速い選手は週60~100km程度の距離を踏むこと、コンディションを整えて練習に参加していること、心拍数を上げずに長く速く走っていることが分かった。

参考文献

- [1] 山内長承, 「Pythonによるデータ解析入門」, オーム社, 2018年, 東京