機械学習による競馬の勝因分析

2017SS096 山崎翔平
指導教員：小市俊悟

1　はじめに

日本の競馬では様々なデータを用いた着順予想がされている．このような予想は基本的に，個人の経験に基づくものであるが，レース結果に関連がありそうなデータを集め，機械学習を用いて分析すれば，同じような予想ができるのではないかと考えた．

そこで，競馬データの収集を行ったが，競馬データでは，比較不可能なカテゴリーデータが多いことに気がついた．対象となるデータがカテゴリーデータの場合，適用できる機械学習の手法も限定的となる．

本研究では，カテゴリーデータに対しても適用できるとされる決定木分析を用いることとし，それによるレース結果，特に 3 着以内に入るか否かの予想を行う．はじめに One-Hot ベクトル形式への変換を利用し，次に貪欲法を用いて，カテゴリーデータを直接扱う手法の提案をする．

2　データについて

競馬のレースで最も有名な有馬記念について，1986 年 – 2019 年の 1 着から 10 着までの馬名，枠順，性別，年齢，騎手，人気のデータを集め [1] さらにその馬の，前走のレース，前走の成績，産駒（父馬），脚質（逃げ/先行/差し/追い込み），前走との馬体重差のデータを集めた [2].

3　決定木分析と One-Hot ベクトル

決定木分析は，判別したい事柄として，本研究では，レースで 3 着内に入るか否かということを指定した．続いて，作成する決定木とは，データに対して適用する条件の集まりである．それらの条件は，木構造で指定される順番に従って適用され，根に相当する条件から木の末端に相当する条件まで適用したときに，判定したい事柄が明らかになっているほど，良い決定木が得られたことになる．

Python の scikit-learn が提供する決定木分析のプログラムは，現在のところ，カテゴリーデータに十分に対応していない．そこで One-Hot ベクトル形式への変換をまずは利用する．その変換では，例えば A, B, C を値として取る属性 X があるとき，A, B, C を新たな属性として設定し直し，属性 A, B, C のそれぞれについて，該当するものを 1，該当しないものを 0 で表す．属性が増えることになるが，これにより，カテゴリーデータは疑似的に数値データとなる．

4　分析結果と考察

図 1 は決定木分析により得られた決定木の左下部分を切り出して表示したものである．最上位のノードに現れる X[198] には，サンデーサイレンス産の駒か否かが 0 と 1 で

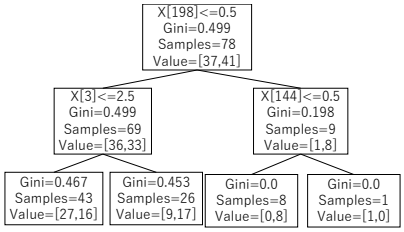

図 1　競馬の決定木の一部

示されているので，“X[198] <= 0.5” という条件により，データは，サンデーサイレンス産駒のデータであれば右側に，そうでなければ左側に振り分けられる．最上位と 2 段目のノードの Samples をそれぞれ見ると，“X[198] <= 0.5” という条件は，78 個のデータに適用され，そのうち，真となり左に分類されたのは 69 個のデータで，偽となり右に分離されたのは，9 個のデータである．最下層の左から 3 つ目のノードには，“Value = [0,8]” と記載されているので，このノードまでの条件の適用結果として，3 着内の馬のデータが 8 個該当し, 3 着外の馬のデータは 1 つも該当しないことがわかる．このような 3 着内と 3 着外のどちらか一方のみのデータが占有する場合，そのノードまでに適用された条件は，判別力が高い条件 (の列) であると言える．100 %の確率で 3 着以内に来る馬の特徴は下記であった．

- 6 番人気以下である．
- 前走の着順が 5 着以内.
- サンデーサイレンス産駒である．
- 前走天皇賞（春）出走していない．

5　問題点

One-Hot ベクトルを採用すると，属性が膨大になり．決定木も複雑になりやすい．カテゴリーデータを直接扱うことの難しさは，主に処理に必要な計算量に起因する．大小比較が不可能な値を取る属性について，取りうる値が n 個あるとき，決定木のノードにその属性を対象に条件を設定することは，実質的に，要素数が n 個の集合から，一つの部分集合を選ぶことに等しい．n が大きい場合にはすべての部分集合を考えることは，計算量的に不可能であるので，可能な方法の一つは，貪欲法により，必ずしも最適とは限らないが，よいと考えられる部分集合を選択することである．6 節で情報量利得を説明したうえで，7 節でこのような考えに基づく方法を提案する．

また，騎手の属性には，現役を引退している騎手も多くいるので，騎手をそれぞれの世代の勝利数などによってラ

ンク分けし，そのランクを利用する．

6　カテゴリー属性による分割

属性 C について，データの集合 D の，各データ d が示す特徴量を $C(d)$ と表す．さらに，$F(C)$ を $F(C) = \{C(d) \mid d \in D\}$ とする．このとき，$F(C)$ の部分集合 $\tilde{F}(C)$ を選べば，データ D を $D_+(\tilde{F}(C)) = \{d \in D \mid C(d) \in \tilde{F}(C)\}$ と $D_-(\tilde{F}(C)) = \{d \in D \mid C(d) \notin \tilde{F}(C)\}$ の 2 つに分割できる．データ D の部分集合 $\tilde{D}$ に対しても，$\tilde{D}_+(C), \tilde{D}_-(C)$ を同様に定める．

決定木分析により説明したい属性 T は，各データ d が $T(d) = 1$ または $T(d) = 0$ を示すとする．このとき，次で定義される数値 (確率) p, p_+, p_- を考える．

$$
\begin{aligned}
p &= \frac{|\{d \in \tilde{D} \mid T(d) = 1\}|}{|\tilde{D}|}, \\
p_+ &= \frac{|\{d \in \tilde{D}_+(\tilde{F}(C)) \mid T(d) = 1\}|}{|\{\tilde{D}_+(\tilde{F}(C))\}|}, \\
p_- &= \frac{|\{d \in \tilde{D}_-(\tilde{F}(C)) \mid T(d) = 1\}|}{|\{\tilde{D}_-(\tilde{F}(C))\}|}.
\end{aligned}
$$

これらを用いて，情報量利得と呼ばれる次を求める．

$$
\begin{aligned}
&IG(\tilde{D}, \tilde{F}(C), T) = \\
&\quad - p \log_2 p - (1-p) \log_2 (1-p) \\
&\quad - \frac{|\tilde{D}_+|}{|\tilde{D}|} \left(-p_+ \log_2 p_+ - (1-p_+) \log_2 (1-p_+)\right) \\
&\quad - \frac{|\tilde{D}_-|}{|\tilde{D}|} \left(-p_- \log_2 p_- - (1-p_-) \log_2 (1-p_-)\right) \quad (1)
\end{aligned}
$$

情報量利得は，基本的に，得られる分割 $\tilde{D}_+, \tilde{D}_-$ がそれぞれ $T(d) = 1$ であるデータか $T(d) = 0$ であるデータに占有されているほど大きくなる．この事実に基づき，情報量利得を 0 以上で最大にするカテゴリー属性 C と $\tilde{F}(C)$ を用いて，データ D を細分していくことで，属性 T を説明しようとすることが，決定木を作成することに対応する．

7　貪欲法による分割決定

データの部分集合 $\tilde{D}$ に対して，情報量利得 (1) を最大にするとは限らないが，カテゴリー属性 C やその特徴量の部分集合 $\tilde{F}(C)$ を求めるために，次のような貪欲法を提案する．カテゴリー属性の種類は多くないものとし，カテゴリー属性それぞれについて，順次，下記の方法に従って特徴量の部分集合を求める．

$\tilde{F}(C)$ を求めるアルゴリズム

0: $\hat{F} = \emptyset$ とし，1 に進む．

1: 各特徴量 $f \in F(C) \setminus \hat{F}$ について，$\hat{F} \cup \{f\}$ とした場合の情報量利得 (1) を求める．求めた情報量利得の最大値が 0 より大きいとき，最大値を与える特徴量を f^* とし，2 へ進む．0 より大きい情報量利得が得られない場合は，$\tilde{F}(C) = \hat{F}$ として終了する．

2: $\hat{F} \leftarrow \hat{F} \cup \{f^*\}$ として，1 に戻る．

8　新しい手法を用いた決定木

7 節で述べた新しい手法を用いて得られた決定木の一部を図 2 に示す．以下の騎手かつ産駒であれば必ず 3 着以内に入る．

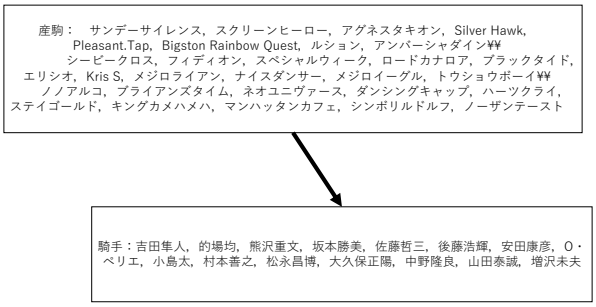

図 2　新しい手法を用いた決定木

騎手をランク分け [3] した場合は次が得られた．

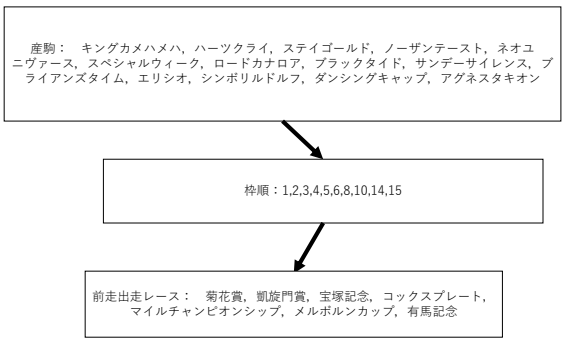

図 3　騎手をランク分けした時の決定木の一部

図 3 にある 3 つの条件をすべて満たすと必ず 3 着以内に入る．得られた決定木には，騎手のランクが用いられなかったので，現役を引退している騎手とも比較しやすくしたものの予想にはあまり重要ではないという結果になった．

9　おわりに

本研究では，ほとんどがカテゴリーデータである競馬のデータに関して，決定木分析を適用する方法を探り，貪欲法を用いて新たな方法を提案した．

参考文献

[1] 有馬記念（過去 G1 成績）JRA（アクセス 2020/6/25）
https://www.jra.go.jp/datafile/seiseki/g1/arima/index.html

[2] 競馬・netkeiba.com（アクセス　2020/6/25）
https://www.netkeiba.com/

[3] 騎手リーディング 競馬データベース-netkeiba.com　(アクセス　2020/10/18)
https://db.netkeiba.com/?pid=jockey_leading/