データマイニング手法を用いた競馬研究

A Study of Horse Racing Using Data Mining Methods

金枝　隆之介†　　華山　宣胤††
Ryunosuke kanaeda, †　　Nobutane Hanayama, ††
†,††尚美学園大学　芸術情報学部　情報表現学科
†,††Department of Information Technology, Shobi University

概要：　競馬については、競走馬に関するものやレースに関するものなど入手可能なデータは膨大であるが、どのデータがレース結果の予想に有用であるかははっきりしていない。そこで、本研究では、データマイニングの手法を用いて競馬データを分析し、勝馬を予測するシステムを構築する。また、構築したシステムによる予想と実際の結果の比較検討を行う。

1．はじめに

競馬データは、競走馬に関するデータ（馬データ）とレースに関するデータ（レースデータ）に大別される．馬データは出走馬の血統、体重の変化、前レースまでの成績などの競走馬に関するものであり、レースデータは馬場の開催地、距離、出走条件、馬場状態などのレースに関するものである．そして、これらのデータに基づいて競馬専門紙やスポーツ新聞などではレース結果の予想が行われている．しかし、それらデータの項目数は膨大であり、どのデータがレース結果の予想に有用であるかははっきりしていない．実際、競馬予想の専門家によって、参考としている項目が異なり、また解釈の方法もまちまちである。つまり、“何から何がわかるか分からないデータベース”といえる．そこで本研究では、膨大な競馬データにデータマイニング手法を適用することにより、統計的な分析結果に基づく勝馬予測システムを構築し、その有用性を検証する．

2．分析手順

提案する分析手順は下記のとおりである．

(1) 競馬データの取得
(2) データの標準化
(3) 標準化されたデータのクラスター分析
(4) クロス集計表の作成
(5) 多群判別分析による新規レースの予想

2.1. 競馬データの取得

使用するデータは、Yahoo！スポーツにより公開されている2007年度のGIデータを使用する．その後必要のないデータの除外を行う．

図表１．馬データ（Excel）

No	レース名	レースクラス	馬名(GI出走馬)	馬番	内枠・外枠(1.0／頭数の半)	性別(牡1・牝0)	年齢	斤量	馬体重	馬体重増減	人気	単勝オッズ
125	ダービー	A	ウオッカ②	3	1	0	3	55	482	-4	3	10.5
126	ダービー	A	アサクサキングス③	16	0	1	3	57	488	-6	14	84.5
127	ダービー	A	アドマイヤオーラ②	14	0	1	3	57	442	-2	4	12.7
128	ダービー	A	サンツェッペリン②	12	0	1	3	57	476	10	9	33
129	ダービー	A	ドリームジャーニー②	8	1	1	3	57	412	2	8	32.8
130	ダービー	A	ゴールデンダリア	4	1	1	3	57	462	2	6	23
131	ダービー	A	フサイチホウオー②	15	0	1	3	57	508	2	1	1.6
132	ダービー	A	ナムラマース②	11	0	1	3	57	456	-6	10	36.3
133	ダービー	A	ヴィクトリー②	17	0	1	3	57	472	-4	2	8.2
134	ダービー	A	フライングアップル②	18	0	1	3	57	498	-6	12	72
135	ダービー	A	タスカータソルテ①	1	1	1	3	57	450	4	7	23.7
136	ダービー	A	マイネルフォーグ②	6	1	1	3	57	440	2	18	200.9
137	ダービー	A	ローレルゲレイロ③	13	0	1	3	57	460	0	13	79.1
138	ダービー	A	トーセンマーチ	5	1	1	3	57	492	2	15	86.7
139	ダービー	A	フィニステール	7	1	1	3	57	510	2	11	53.2
140	ダービー	A	ヒラボクロイヤル①	9	0	1	3	57	488	6	5	21.1
141	ダービー	A	ゴールドアグリ②	2	1	1	3	57	478	-6	16	135.8
142	ダービー	A	プラテアード	10	0	1	3	57	500	-2	17	152.7

図表２．レースデータ（Excel）

レース名	中京	阪神	中山	京都	東京	右・左回り(1.0)	距離	混合・牡・牝(0.1.2)	2歳戦・3歳戦・古馬戦(0.1.2)	国際(1.0)	天気(晴れ1・その他0)	馬場状態(良1・その他0)	出走頭数
高松宮記念	1	0	0	0	0	0	1200	0	2	1	1	0	18
桜花賞	0	1	0	0	0	1	1600	2	1	0	1	1	18
皐月賞	0	0	1	0	0	1	2000	0	1	0	1	1	18
天皇賞(春)	0	0	0	1	0	1	3200	0	2	1	1	1	16
NHKマイルカップ	0	0	0	0	1	0	1600	0	1	0	0	0	18
ヴィクトリアマイル	0	0	0	0	1	0	1600	2	2	1	1	1	18
優駿牝馬(オークス)	0	0	0	0	1	0	2400	2	1	0	1	1	18
東京優駿(日本ダービー)	0	0	0	0	1	0	2400	0	1	0	1	1	18
安田記念	0	0	0	0	1	0	1600	0	2	1	1	1	18
宝塚記念	0	1	0	0	0	1	2200	0	2	1	0	0	18
スプリンターズステークス	0	0	1	0	0	1	1200	0	2	1	0	0	16
秋華賞	0	0	0	1	0	1	2000	2	2	0	1	1	18
菊花賞	0	0	0	1	0	1	3000	0	1	0	1	1	18
天皇賞(秋)	0	0	0	0	1	0	2000	0	2	1	1	0	16
エリザベス女王杯	0	0	0	1	0	1	2200	2	2	1	1	1	13
マイルチャンピオンシップ	0	0	0	1	0	1	1600	0	2	1	0	1	18
ジャパンカップ	0	0	0	0	1	0	2400	0	2	1	1	1	18
有馬記念	0	0	1	0	0	1	2500	0	2	1	1	0	15

2.2. データの標準化

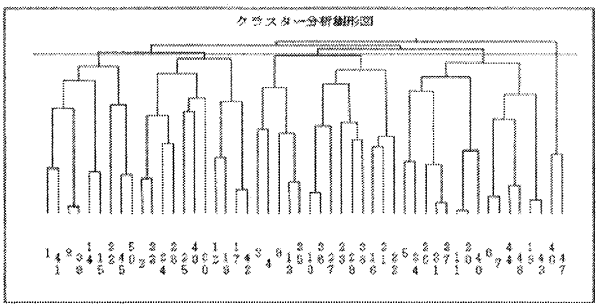
競馬データには様々なタイプや単位の項目が含まれるため，下記の計算式によりデータの標準化を行う．

標準化データ
＝（個々のデーター平均値）／標準偏差

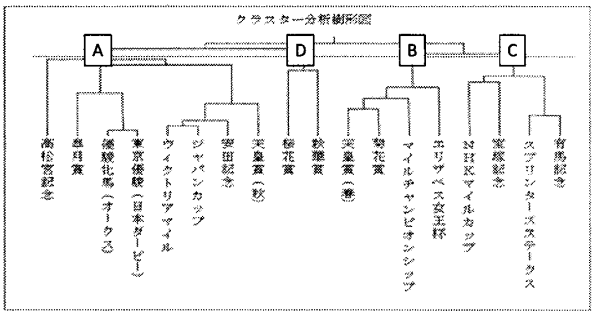
2.3. 標準化されたデータのクラスター分析

標準化された馬データ、レースデータのクラスター分析を行う．

図表３．馬データのデンドログラム



図表４．レースデータのデンドログラム



2.4. クロス集計表の作成

クラスター分析の結果に基づいて「レースクラスター」×「馬クラスター」別の勝率算出する（図表５）．図表５は，各馬クラスターに属する馬が、各レースクラスターに出走した場合の勝率を表している．

図表５．2007年データに基づいた馬クラスター別の勝利率

馬分類 / レース分類	1	2	3	4	5	総計
A	17%	6%	0%	22%	0%	44%
B	6%	0%	0%	17%	0%	22%
C	0%	6%	11%	6%	0%	22%
D	0%	0%	0%	11%	0%	11%
総計	22%	11%	11%	56%	0%	100%

2.5. 多群判別分析による新規レースの予想

多群判別分析を用いて，新規レース（2008年ジャパンカップ）に出走した馬を図表５のA～Dの馬クラスターに分類（判別）する（図表６）．

図表６．2008年のジャパンカップに出走した各場の判別結果と実際の順位（結果）

着順	馬名（G1出走馬）	属するクラスター
1	スクリーンヒーロー	1
2	ディープスカイ	5
3	ウオッカ	5
4	マツリダゴッホ	1
5	オウケンブルースリ	1
6	メイショウサムソン	4
7	ネヴァブション	2
8	アサクサキングス	1
9	パープルムーン	1
10	トーホウアラン	1
11	オースミグラスワン	5
12	アドマイヤモナーク	5
13	シックスティーズアイコン	1
14	ペイパルブル	1
15	トーセンキャプテン	1
16	ダイワワイルドボア	1
17	コスモバルク	5

3．分析結果の検討

図表５と６から，2008年のジャパンカップの勝ち馬は，「馬クラスター４」に属するメイショウサムソンと予想されたが，実際は「馬クラスター１」に属するスクリーンヒーローであった．しかし，図表５から「馬クラスター１」に属する馬も「馬クラスター４」に匹敵する勝率であることは注目すべきであろう．また，勝率が「馬クラスター５」に属する２頭が上位に位置しているが，ディープスカイは「馬クラスター１および４」に近く，ウォッカは「馬クラスター４」に近かったことも注目すべきであろう．

4．まとめ

本研究では，勝ち馬予想手順を提案したが，結果は実際の勝ち馬予想に有用な段階に達しなかった．今後は用いる項目の選択方法を検討するなどの改善を行う予定である．

参考文献

[1] 月本 洋，実践データマイニング—金融・競馬予測の科学，オーム，1999

[2] 小岩井 弥，競馬データを120%活用するクォンツ分析理論，毎日コミュニケーションズ，1996

[3] 竹内 光悦，酒折 文武，Excelで学ぶ理論と技術 多変量解析入門，ソフトバンククリエイティブ，2006