利用贝叶斯判别函数预测网球比赛胜负的可行性分析

赵得霖、曾思栋

【摘要】

本文旨在利用贝叶斯判别分析的相关知识，分析在一种特定的赌球玩法中，庄家利用即时比赛数据提升比赛结果预测的精准度，并以此诱盘谋取暴利的可行性。

【关键词】

相关系数　判别分析　贝叶斯判别函数　正确率

目录

引言

赌球，是指人们拿足球、网球、篮球、橄榄球等比赛有关的客观事实进行赌博的行为。它利用的是人们想要一夜暴富的心理，虽已属违法行为，却屡禁不止，甚至还扩展出了许多不同的玩法。本课题将以一种在球赛开始后的一段时间内仍可下注的赌球玩法为例，并以网球比赛为研究对象，简要分析为什么“只要下注，等待赌客的往往是必输”。

首先，庄家制定的每一个赔率，都建立在赌博集团超强的资讯能力、庞大的精算师和数学家团队的缜密分析之上。所以我们猜测，在该种赌球玩法下，庄家可以收集到选手的即时技术数据进行分析，并基于以往的比赛数据预测比赛结果。那么，一方面，赌客缺少全面的即时比赛数据；另一方面，赌客也难以收集到大量的历史比赛数据。这种信息的不对称使庄家能够不断通过调整赔率来诱导赌客，以确保自己“稳赚不赔”。

上述的思路其实就是通过即时数据将先验概率转化为后验概率，这与贝叶斯判别分析的基本思想——将待判样品配属给后验概率最大的总体是相吻合的。所以接下来，我们将验证贝叶斯判别分析是否能利用比赛开始后一段时间内的即时技术数据，帮助庄家提升比赛结果预测的准确性。

一、数据

1. 数据来源：http://www.tennisabstract.com/
2. 数据筛选：
    1) 根据 2008 年-2018 年的世界排名，选定 10 位世界排名长时间位于前十的男选手，获取他们十年间的所有在案对局数据；
    2) 从中筛选出世界排名前二十选手之间的对局数据，以平衡胜负场次的比例；
    3) 剔除数据缺失的比赛：部分比赛数据不完整，影响整体的数据处理；
    4) 剔除不完整的比赛：如选手中途退赛的比赛或由于特殊情况终止的比赛等。
3. 数据处理：

1） 删除无关数据，如：比赛日期、赛事名称、轮次等；
2） 以数值型数据代替非数值型数据，方便后续处理，如：数字“1“代表胜利，数字”2“代表败北；
3） 由于不同比赛的总盘数存在差异，故将原始数据中与“破发”相关的四类数据转化为盘均数据，如：盘均挽救破发点数=挽救破发点总数/本场比赛总盘数。

4. 数据科学性分析：
1） 数据来源于专业的网球数据分析网站，可信度高；
2） 剔除可能造成负面影响的比赛，提高数据分析的准确性；
3） 删除无关数据，使数据保持精简、实用，避免无关影响。

二、相关性分析

为建立判断比赛胜负的判别函数，我们选取了十三个可能对比赛结果有影响的指标，但为保证所建立判别函数的稳定性和预测的准确性，我们应首先对这十三个指标间的相关系数进行计算(cor.R)。

	1	2	3	4	5	6	7	8	9	10	11	12	13
DR	1												
ACE	0.291	1											
vACE	-0.16	0.026	1										
DF	-0.2	-0.01	0.046	1									
X1stIn	0.215	-0.05	0.058	-0.3	1								
X1st	0.657	0.492	0.048	-0.05	-0.02	1							
X2nd	0.601	0.138	0.033	-0.32	0.14	0.322	1						
v1st	0.492	0.016	-0.52	-0.07	0.11	0.159	0.143	1					
v2nd	0.482	0.008	-0.12	-0.06	0.134	0.193	0.186	0.304	1				
aBPsvd	-0.42	-0.23	-0.01	0.097	-0.08	-0.42	-0.34	-0.09	-0.12	1			
aBPfsvd	-0.66	-0.34	-0.06	0.19	-0.2	-0.75	-0.64	-0.18	-0.24	0.386	1		
aBPcnv	0.535	-0.01	-0.38	-0.07	0.135	0.182	0.179	0.751	0.641	-0.07	-0.19	1	
aBPfcnv	0.203	-0.01	-0.18	-0.07	0.058	0.073	0.048	0.325	0.266	0.004	-0.04	0.324	1

（表一）

数据说明：

DR	接发球局得分率/发球局失分率
ACE	发球 ACE 率
vACE	对手发球 Ace 率
DF	双发失误率

X1stIn	一发成功率
X1st	一发得分率
v1st	对手一发得分率
X2nd	二发得分率
v2nd	对手二发得分率
aBPsvd	盘均挽救破发点数
aBPfsvd	盘均被破发数
aBPcnv	盘均破发数
aBPfcnv	盘均未成功破发数

（表二）

从表一我们可以看出：

每两个指标之间的相关系数均介于0-0.7之间,整体而言属于中低程度相关，各指标相互之间的影响较小，适合进行判别分析。（一般来说，线性相关高低可按三级划分：|r|＜0.4 为低度线性相关；0.4≤|r|<0.7 为中度相关；0.7≤|r|＜1 为高度线性相关）

三、基于样本全体建立总体判别函数

利用全体数据(data_20vs20.csv）训练总体判别函数(Function.B.R)并验证总体判别函数(Compare.F.R)的具体步骤:

1) 随机选取300场胜场、300场负场进行训练,得出贝叶斯判别函数:Y_1(胜利)、Y_2(败北)；
2) 另外随机选取 200 场比赛的数据代入函数 Y_1，Y_2进行验证。若 Y_1＞Y_2，则将该场比赛判为胜；若 Y_1＜Y_2，则将该场比赛判为负；
3) 将判断出的情况与实际的胜负比较，计算出判别函数判别的准确率。

其中一次的判别函数为：

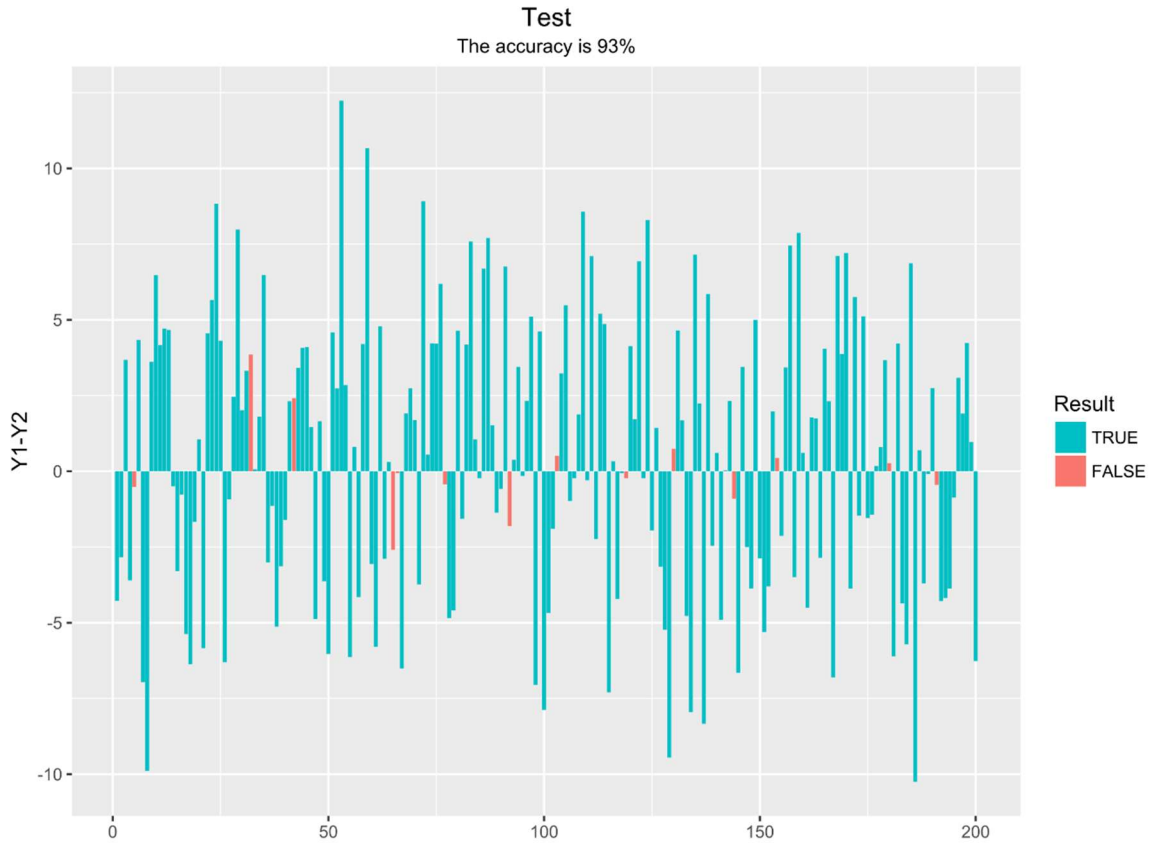
	Y_1	Y_2
DR	-59.990958	-62.147137
ACE	-72.935853	-70.994820
vACE	-32.914845	-30.009758
DF	285.605405	271.882775
X1stIn	261.609633	261.513080

X1st	625.427066	623.701042
X2nd	324.528891	321.245822
v1st	181.839356	183.756287
v2nd	164.736364	163.886043
aBPsvd	7.130132	7.002564
aBPfsvd	71.931611	75.077177
aBPcnv	-19.033252	-22.030478
aBPfcnv	-1.447830	-1.677066
Constant	-455.964779	-450.100892

（表三）

Y_1列为 Y_1函数对应变量的系数；Y_2列为 Y_2函数对应变量的系数；最后一行为常数。

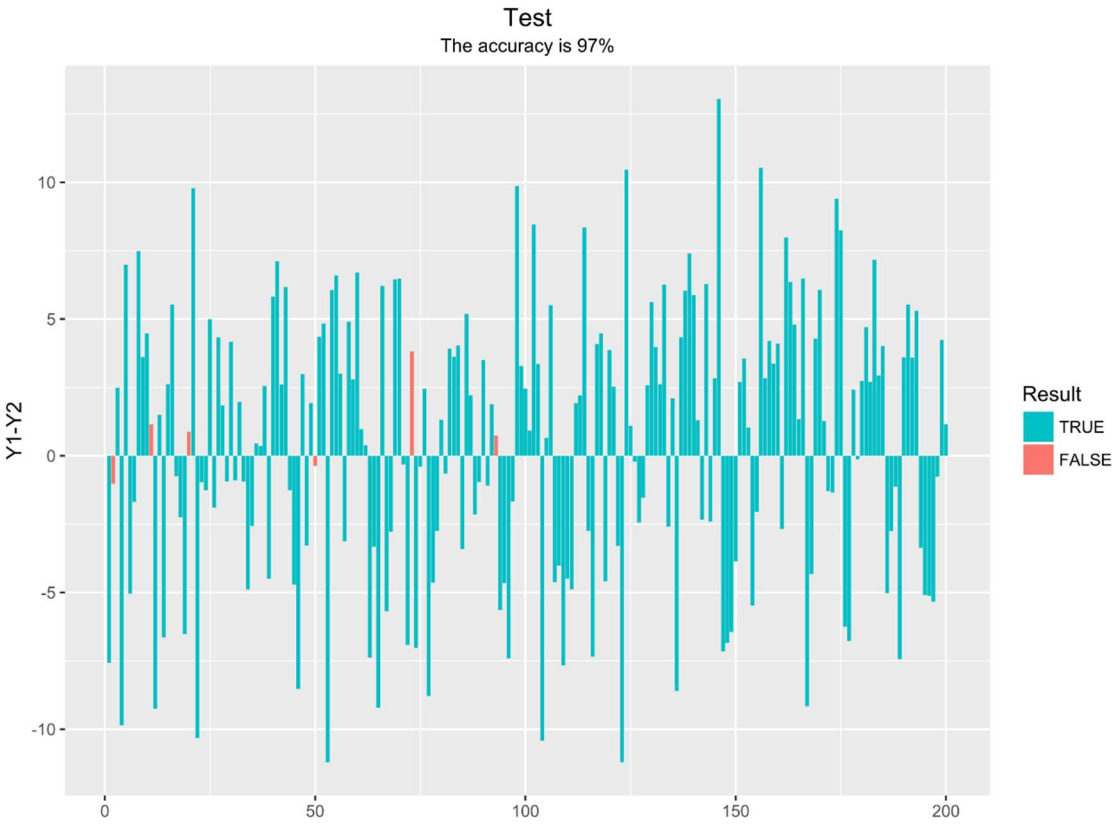
对该判别函数的单次检验结果如下图：

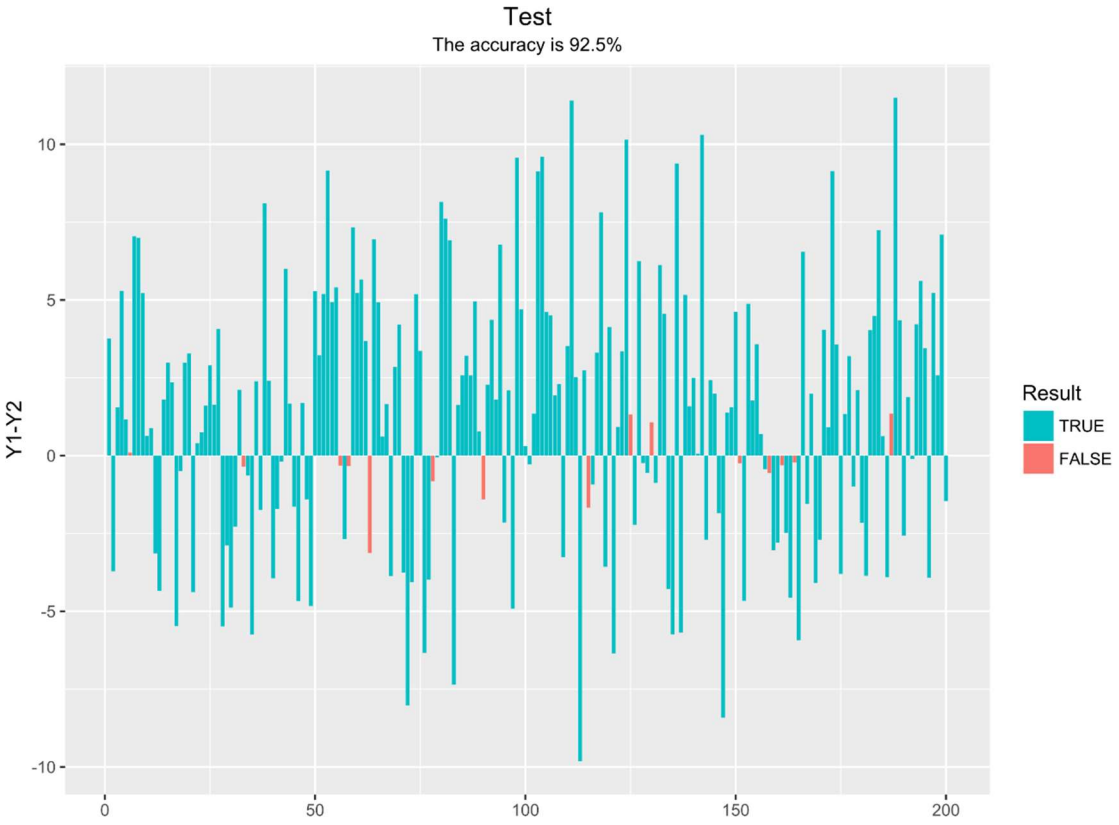


（图一）

位于 x 轴上方的柱形代表该次比赛被判为“胜”，位于 x 轴下方的柱形代表该次比赛被判为“负”；判别正确的柱形被填充上蓝色，判别错误的柱形被填充上红色；本次判别正确率为 93%。

图二、三是重复上述步骤得到的另外两次检验结果：

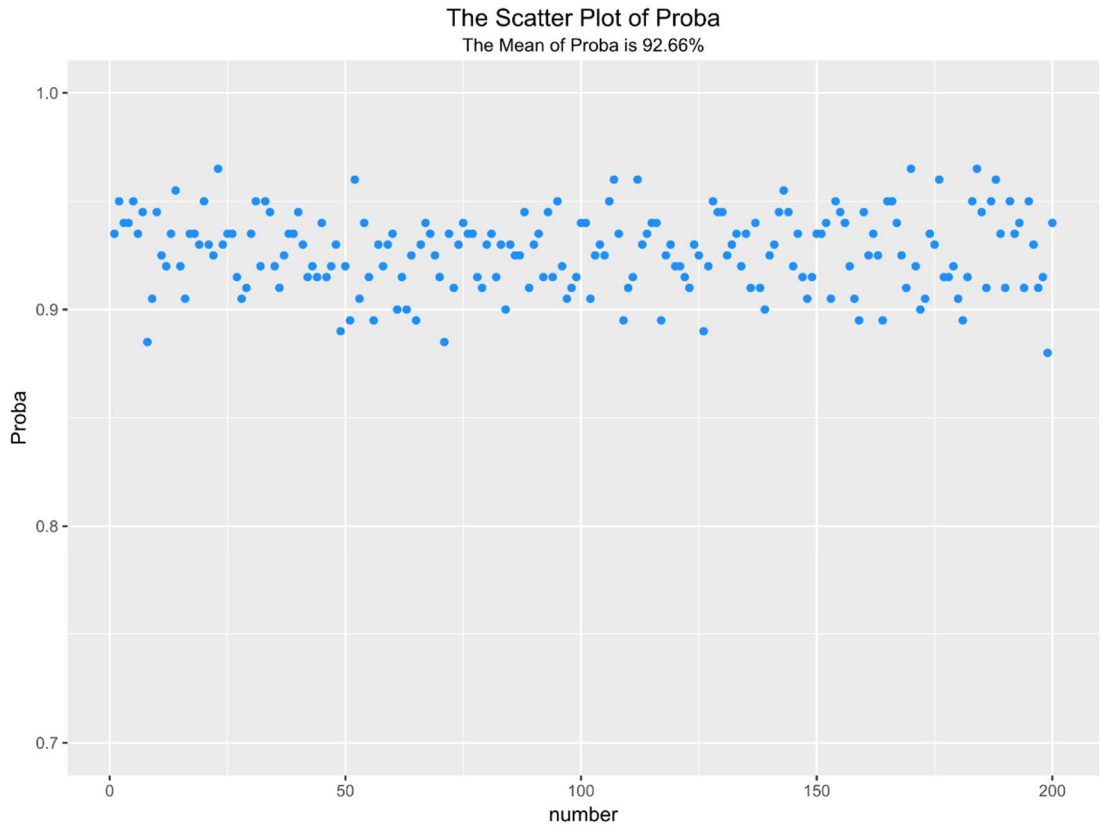




（图二、三）

由以上各图可见，检验的准确率均在 90%以上，且判断错误的场次均为（Y_1-Y_2）较小的场次。

重复上述步骤 200 次，画出正确率散点图（图四）（Proba.M.R），并得出平均正确率：



（图四）

可以看出，判别正确率大多介于 90%-95%之间，最低值大于 87.5%，且平均正确率为 92.66%，准确率高且相对稳定，说明用上述 13 个指标来建立判别函数预测比赛结果是合理的。

四、基于个体样本建立个人判别函数

逻辑上看，通过个体样本建立的个人判别函数比起总体判别函数理应更能准确预测某一特定运动员的比赛结果。接下来我们将基于个体样本建立个人判别函数，并针对相同数量的检验样本，比较个人判别函数与总体判别函数的判别正确率。

我们首先针对特定的运动员：费德勒和德约科维奇的数据分别训练判别函数（随机选取 50 场胜场和 50 场负场），并进行验证（随机选取 200 场比赛）：

两位运动员各项指标之间的相关系数矩阵：

	1	2	3	4	5	6	7	8	9	10	11	12	13
DR	1												
ACE	0.301	1											
vACE	-0.07	0.1	1										
DF	-0.11	-0.07	-0	1									
X1stIn	0.14	0.262	0.091	-0.23	1								
X1st	0.693	0.459	0.164	-0.03	-0.02	1							
X2nd	0.66	0.136	0.091	-0.14	0.107	0.373	1						
v1st	0.543	0.011	-0.49	0.014	-0.06	0.181	0.174	1					
v2nd	0.468	-0	-0.01	0.056	0.032	0.184	0.15	0.196	1				
aBPsvd	-0.5	-0.27	-0.14	0.084	-0.17	-0.47	-0.46	-0.11	-0.13	1			
aBPfsvd	-0.65	-0.35	-0.19	0.071	-0.19	-0.74	-0.66	-0.15	-0.19	0.454	1		
aBPcnv	0.568	0.018	-0.3	0.091	-0.07	0.251	0.16	0.71	0.587	-0.11	-0.14	1	
aBPfcnv	0.105	0.052	-0.11	-0.04	0.012	0.007	-0.02	0.255	0.155	0.059	0.032	0.216	1

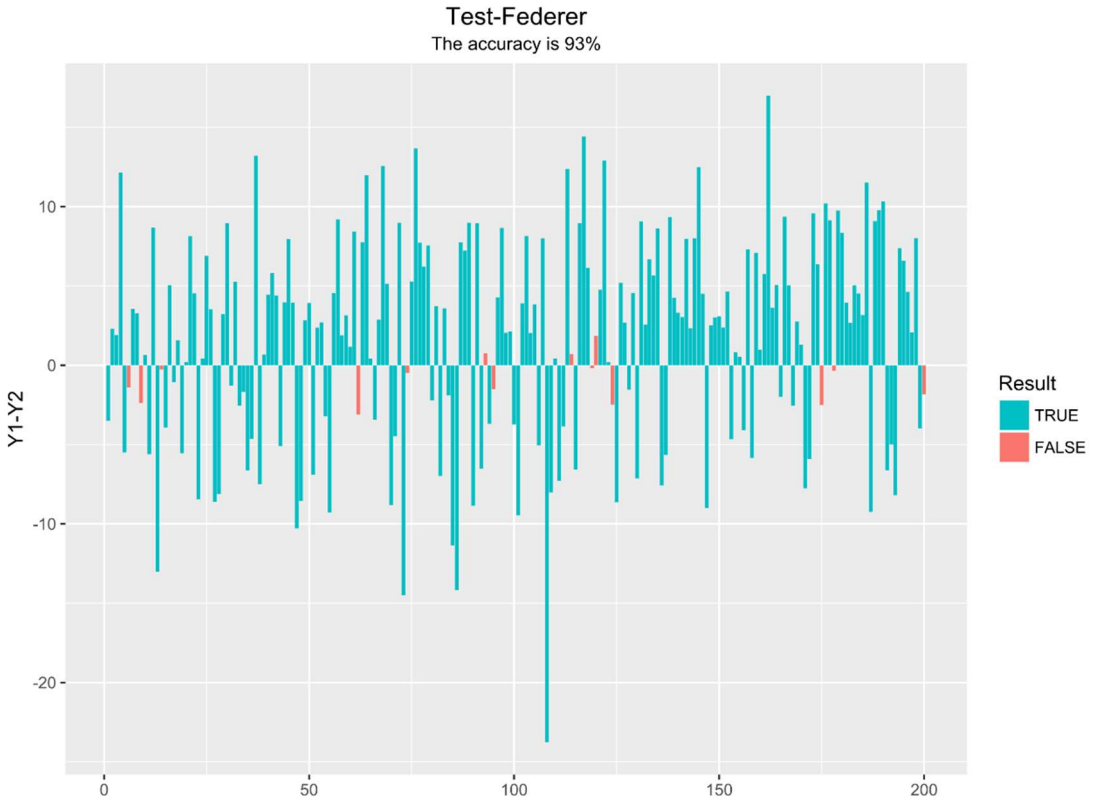
（费德勒，表四）

	1	2	3	4	5	6	7	8	9	10	11	12	13
DR	1												
ACE	0.38	1											
vACE	-0.21	0.042	1										
DF	-0.24	-0.14	-0.06	1									
X1stIn	0.239	0.201	0.062	-0.32	1								
X1st	0.669	0.497	-0.03	-0.05	0.06	1							
X2nd	0.515	0.147	0.091	-0.41	0.085	0.15	1						
v1st	0.487	0.043	-0.5	-0.01	0.081	0.19	-0	1					
v2nd	0.583	0.055	-0.14	-0.12	0.139	0.268	0.165	0.243	1				
aBPsvd	-0.41	-0.25	0.085	0.136	-0.08	-0.38	-0.32	-0.17	-0.11	1			
aBPfsvd	-0.63	-0.38	-0.06	0.225	-0.21	-0.69	-0.53	-0.16	-0.31	0.273	1		
aBPcnv	0.591	0.021	-0.42	-0.05	0.094	0.224	0.073	0.776	0.622	-0.12	-0.18	1	
aBPfcnv	0.173	-0.01	-0.14	-0.13	-0.04	0.086	0.054	0.225	0.246	-0.02	-0.03	0.256	1

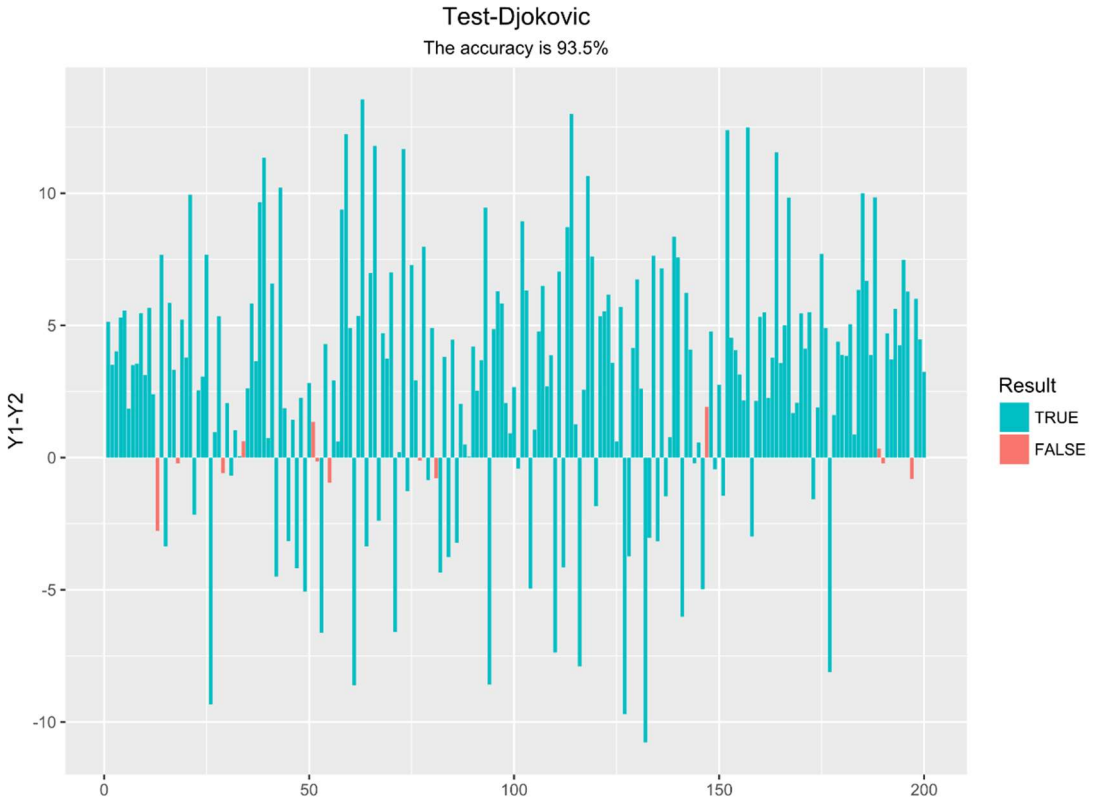
（德约科维奇，表五）

对于这两名运动员来说，每两个指标之间的相关系数基本都介于 0-0.7 之间，整体而言属于中低程度相关，各指标相互之间的影响较小，均适合进行判别分析。

判别结果及对应的检验结果（以下情形均为某一次随机出现的结果）：

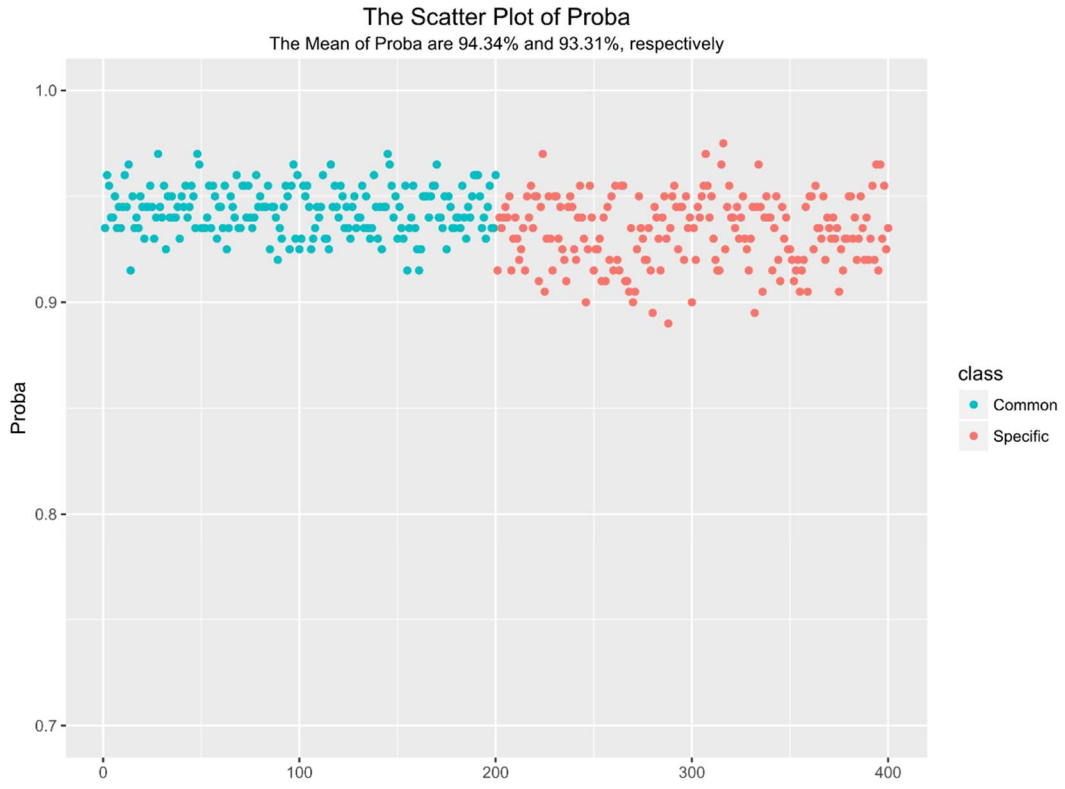


（费德勒，图五）

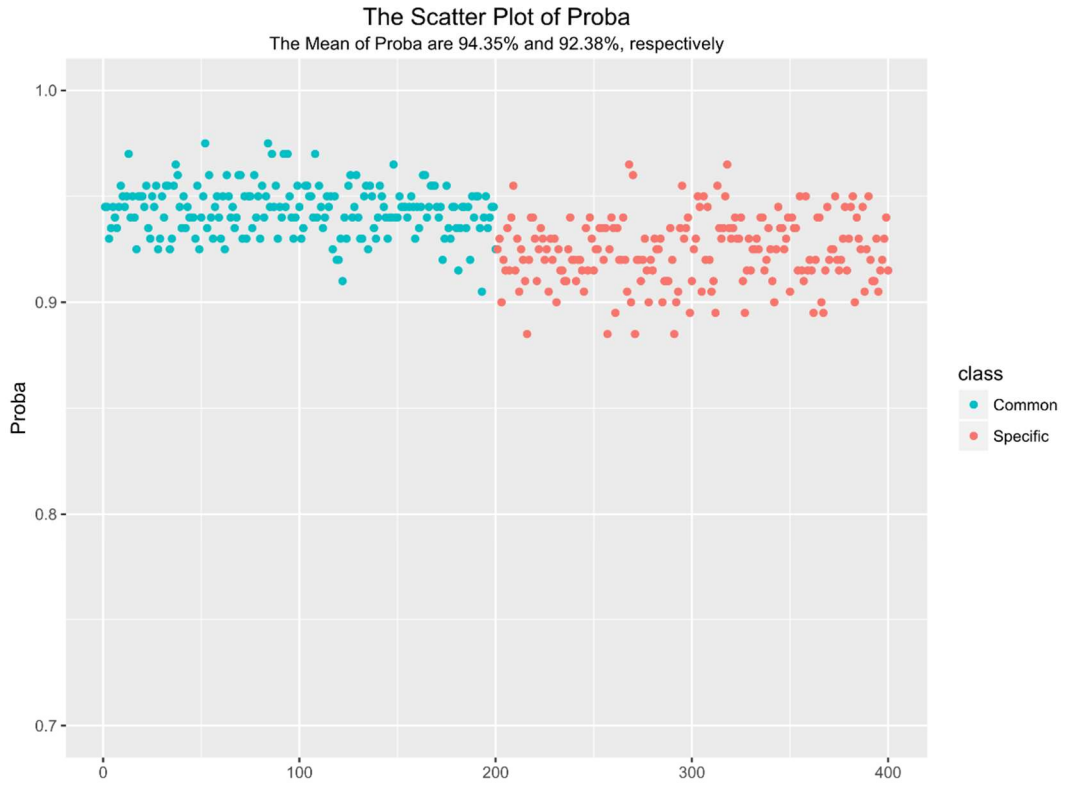


（德约科维奇，图六）

判别正确率散点图和平均正确率(Compare.P.M.R)（左半图对应总体判别函数，右半图对应个人判别函数）：



（费德勒，图七）



（德约科维奇，图八）

我们发现，总体判别函数的正确率不仅比个人判别函数更稳定，且平均值更高，这可能与建立个人判别函数时训练样本量较少有关。因此，接下来我们将使用总体判别函数来预测特定运动员的比赛结果。

五、即时数据预测比赛结果的可行性（以费德勒为例）

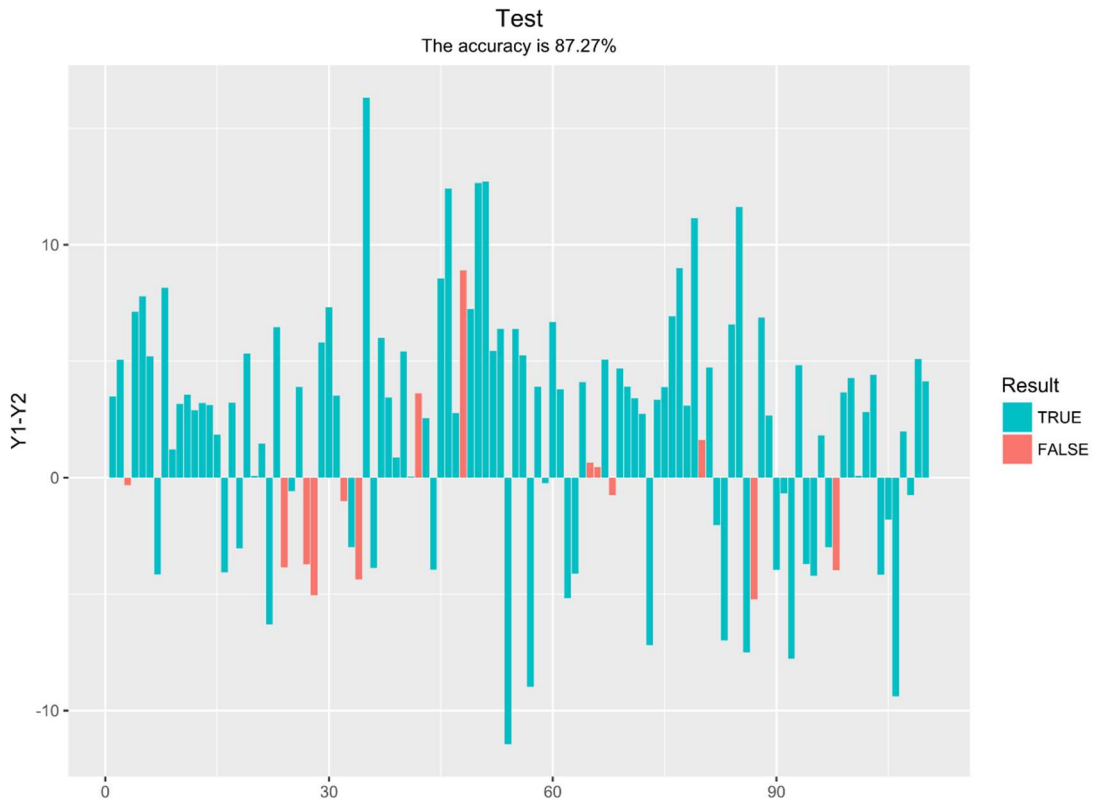
根据上述结论，我们通过多次生成总体判别函数，比较各判别函数判断费德勒比赛结果的正确率，选出正确率最高（97.5%）的判别函数如下：

	Y1	Y2
DR	-64.124439	-66.343320
ACE	-77.551012	-76.655098
vACE	-15.963404	-11.152308
DF	433.566752	421.168699
X1stIn	299.864225	298.936727
X1st	623.934904	622.973107
X2nd	376.768081	374.498550
v1st	190.920091	190.900220
v2nd	155.208687	153.432766
aBPsvd	7.619324	7.579242
aBPfsvd	72.968492	76.129460
aBPcnv	-19.280594	-21.913192
aBPfcnv	-1.883177	-1.993420
Constant	-481.390128	-475.765810

（表六）

为了方便数据采集，我们将第一盘比赛的技术数据作为仍可下注的时间段中某一时间点的即时比赛数据，并选出了费德勒 100 余场比赛的第一盘数据（data_Federer_first_match.csv）作为检验样本，其中第一盘比赛的胜负不在考虑范围之内。

由上述判别函数得出判别结果如下：



（图九）

利用上述判别函数，通过第一盘的比赛数据预测最终比赛结果的正确率达到了 87.27%，显著高于 75%。这说明通过贝叶斯判别分析，用即时数据预测比赛结果是可行的。

六、结论

通过以上分析可以得出结论：庄家确实可以通过分析即时数据提升对比赛结果预测的准确性，从而诱盘谋取利益，与此相对的，底层赌客久赌必输。赌球既不公平也不合法，所以我们应远离赌球，谨防赌球陷阱。