分类号 O21 密级 公开
UDC 编号

雲南大學

硕士研究生学位论文

题 目 网球比赛的赛果预测及球员分析
Title Tennis match result forecast and player analysis

学院（所、中心） 数学与统计学院
专业名称 应用统计专业硕士
研究方向 应用统计
研究生姓名 张蓉 学号 12019202033
导师姓名 陈黎 职称 副教授

2022 年 5 月

摘要

随着我国运动员在国际网球赛场上不断取得突破，网球在我国受到越来越多的球迷喜爱，网球博彩市场的发展也日益繁荣，球迷及博彩公司都对网球比赛的赛果预测十分感兴趣。本文通过分析网球比赛影响因素及对网球比赛进行赛果预测，为球迷以及博彩公司提供理论依据和参考价值。

首先，本文研究了网球比赛的影响因素。从发球、接发球以及得分三个环节考虑，选择十六项技术指标构建技术指标体系。采用因子分析法，提取出了影响球员表现的三个公因子，分别为发球局表现因子、接发球局表现因子以及发球稳定性因子。依据三个公因子对世界排名前十六的球员进行评分，将他们按照实力分为三档并进行评价。

其次，本文使用基于排名系统的马尔可夫模型对网球赛果进行预测。根据选手发球得分的概率，递归地计算出选手赢得局、抢七局、盘和比赛胜利的概率。对排名系统进行改进，使用改进的排名系统估计球员发球得分的概率，并将不同类型的场地影响以及发球接发球环节的内在差异纳入模型，对排名系统及模型作出修正。改进的马尔可夫模型正确预测百分比的均值为 0.68822，比起初始马尔可夫模型提高了 0.65%，偏差也有所下降。

接下来，使用机器学习中的支持向量机模型、逻辑回归模型、以及多层感知机模型对网球比赛胜负进行预测，其中表现最好的模型是支持向量机模型，它在测试集上的预测正确率达到 73%。对各个特征进行消融研究，得出决定比赛结果的最重要的特征是球员的历史交手记录。最后进行网球投注，使用 XGboost 模型和赔率数据对所选比赛进行投注，通过对 35%的比赛下注，平均投资回报率为 20%。

本文的创新点在于将马尔可夫模型与排名系统结合，形成了一种新的网球建模方法。论文的研究成果不仅可以对球员提供建议，还对球迷及博彩公司有一定的借鉴意义。

关键词：网球比赛；马尔可夫模型；排名系统；机器学习；网球投注

Abstract

With the continuous breakthrough of Chinese athletes in the international tennis court, tennis is favored by more and more fans in Our country, and the development of the tennis betting market is increasingly prosperous. Both fans and betting companies are very interested in predicting the results of tennis matches. This paper provides theoretical basis and reference value for fans and betting companies by analyzing the influencing factors of tennis matches and predicting the results of tennis matches.

Firstly, this thesis studies the influencing factors of tennis match. Considering the three links of serving, receiving and scoring, sixteen technical indexes are selected to construct the technical index system. Three common factors affecting player performance are extracted by factor analysis method, which are service performance factor, service return performance factor and service stability factor. The top 16 players in the world are rated on three common factors, divided into three grades and evaluated according to their strength.

Secondly, markov model based on ranking system is used to predict tennis match results. The probability of a player winning a set, tiebreak, set and match is recursively calculated based on the probability of a player scoring a serve. The ranking system is improved, and the probability of serving points is estimated by using the improved ranking system, and the influence of different types of courts and the internal differences of serving and receiving links are incorporated into the model, and the ranking system and model are revised. The average prediction percentage of the improved Markov model is 0.68822, which is 0.65% higher than that of the original Markov model, and the deviation is also reduced.

Then, the support vector machine model, logistic regression model and multi-layer perceptron model in machine learning are used to predict the outcome of tennis matches. The best model is the support vector machine model, which has a prediction accuracy of 73% in the test set. The ablation study of each feature shows that the most important feature in determining the outcome of a match is the player's historical record. Finally, tennis betting. Using XGBoost models and odds data to place bets on selected races, the average return on investment was 20% by placing bets on 35% of the races.

The innovation of this paper lies in the combination of Markov model and ranking system, forming a new tennis modeling method. The research results of this paper can not only provide suggestions for players, but also have certain reference significance for

football fans and betting companies.

Key words: Tennis match; Markov model; Ranking system; Machine learning; Tennis betting

目　录

第一章　绪论

1.1 研究背景

网球现在已经发展成为世界上最具有观赏性和最受球迷喜欢的运动之一。职业网球协会每年会在超过 30 个国家举办 60 多站职业网球巡回比赛，作为网球比赛最高水平的四大满贯赛事，澳网、法网、温网、美网尤其受到广大球迷群体的喜爱。2011 年 6 月 4 日，中国选手李娜在法国网球公开赛夺得冠军，这使得她成为中国乃至亚洲历史上第一个夺得大满贯比赛单打冠军的选手，当晚有超过 1 亿中国观众观看了这场决赛。2013 年，安迪穆雷在温布尔登网球公开赛决赛中历史性地击败诺瓦克德约科维奇，这场比赛成为英国全年收视率最高的电视转播，观众人数达到 1730 万。费德勒、纳达尔、德约科维奇、威廉姆斯姐妹、李娜等著名网球运动员不仅成为了越来越多网球爱好者的偶像，也吸引了一大批专家学者以他们为主要研究对象对网球运动进行深入研究。

我国网球运动普及起步较晚，在克服了起点低、群众基础差、国际交流比赛机会少等重重困难后，逐步发展，1998 年才提出了网球职业化的概念。但是由于当时我国承办的各类职业网球赛事级别不高，国内职业网球运动员与国际高水平运动员缺乏交流，水平差距较大，所以网球职业化也仅仅留存于形式。2002 年釜山亚运会中国网球正值新老交替时期，七个项目无一进入四强，中国网球遭遇滑铁卢，迎来黎明前最黑暗的时刻。之后中国网球管理中心明确提出，中国网球要想发展就必须要走网球职业化道路的指导思想[1]。针对国外网球双打选手需要临时组队的情况，制定了以女子双打为突破口的战略方针[2]。2004 年李婷和孙甜甜夺得雅典奥运会女子双打冠军、2006 年郑洁和晏紫先后获得澳网和温网的女双冠军，而李娜的战绩更为显著[3]。她先后获得了北京奥运会单打第四名、2008 年黄金海岸赛冠军、2011 年悉尼公开赛冠军、2011 年澳网亚军（亚洲首个打入大满贯决赛的选手）、2011 年法网冠军（亚洲首个获得大满贯冠军的选手），被国内球迷亲切地称为“亚洲一姐”[4]。李娜在国际赛场上收获的成功使得国人对网球的热情空前的高涨，中国女子网球成为了世界女子网球中一股强大不可忽视的力量。统计 2000-2019 年所有的 ATP 男子网球赛事和 WTA 女子网球赛事得到图 1.1，由图 1.1 可以看出，代表中国出战过国际网球赛事的男子球员和女子球员共计 51 位，这一

数字位居世界第七，这也从侧面反映出了我国网球的飞速发展。

不同于女子网球的迅速发展，我国男子网球发展缓慢，与世界顶级水平之间有着巨大的差距，极少能够看到男子网球运动员在一些国际顶级网球赛事中展露头角，和我国的许多其他的体育项目一样呈现出阴盛阳衰的局面。2020 年，年仅 25 岁的张之臻在 ATP 官方网站的排名突破性地达到了 136 名，这也是中国男子网球选手在 ATP 官方网站排名的历史新高。不同于女子球员在大满贯赛场勇夺桂冠，男子球员至今仍在为进入大满贯的正赛而苦苦努力，只要有人获得一场大满贯的正赛胜利就可以创造中国男子网球的历史了。

近年来，随着网球运动的普及，以及在线体育博彩市场的扩张，网球投注量大幅增加。前面提到的穆雷与德约科维奇的温布尔登决赛，在全球最大的博彩交易所 Betfair 上的投注总额达到了 4800 万英镑，网球投注在网球市场发达的美国及欧洲获得了明显的效益。网球投注作为体育博彩的一个重要组成部分，满足了网球球迷们参与比赛的愿望。现在出现了许多的体育博彩公司，这些公司为球迷进行网球方面的投注提供了依据。国外的主要有 Bet365,Betway,BetVictor 等公司，国内的主要有明升，申博，同乐城等，这使得网球比赛的赛果预测成为了一个全新的挑战[5]。潜在的利益以及科研学者在学术方面的兴趣，推动了对网球比赛精确预测算法的研究。

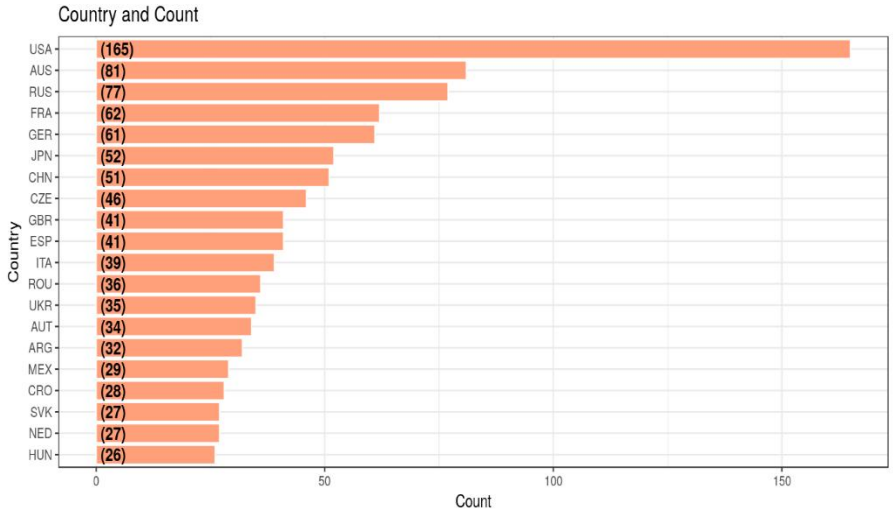

图 1.1 代表各国出战的球员数量

1.2 研究意义

目前，体育竞彩方面多玩的是篮球和足球，网球竞彩玩的不是很多。在理论研究方面，针对篮球和足球的因素分析和赛果预测的模型和方法已经非常成熟，而针对网球的因素分析和赛果预测研究相比较篮球和足球要晚很多，现有的文献多是 2000 年之后的，国内的研究相较国外则更晚。

已有的网球赛果预测模型和方法中，马尔可夫方法比较成熟和体系化，马尔可夫方法中估计选手在自己发球局得分的概率是重要的，确定了选手发球得分后，在比赛中可以随着比赛的进行逐次更新，再进行赛果预测，因此对选手发球得分的概率的估计显得至关重要。

本文尝试使用一种新的网球建模方法，同时利用了排名系统和马尔可夫模型。排名系统用于比较网球运动员的发球评分等级和接发球评分等级。改进排名系统，估计球员在发球局中赢得一分的概率。然后将这些概率值输入马尔可夫链模型，期望得到很好的预测结果。此外，国内发表的网球相关的论文，运用机器学习算法对网球比赛赛果进行预测的很少，网球投注方面的研究更是少之又少，通过本文的研究，希望为国内的网球爱好者以及网球相关的预测分析提供一些理论依据和参考价值。

1.3 国内外研究现状

1.3.1 国内研究现状

国内对于职业网球的研究起步较晚，所以本文分析整理的国内有关职业网球比赛相关的文献都源自 2000 年后，共检索到相关文献 2214 篇。早期相关的文献研究的问题主要集中于中国近代网球的发展、中国网球职业化道路探索等方面[6]。直到 2004 年雅典奥运会，李婷和孙甜甜获得了女子双打金牌，我国女子网球选手在国际网球赛场取得了突破，国内的专家学者这才逐渐开始把研究的方向转向网球比赛影响因素相关的内容[7][8]。2008 年 12 月，中国网球管理中心做出允许球员单飞的决定，对我国网球职业化道路产生了重大而深远的影响，中国网球也正式进入高速发展阶段，以李娜为代表的优秀的网球运动员甚至在大满贯赛事中折冠，备受国际社会瞩目[9]。自此，网球比赛影响因素相关的研究在网球学术领域成为主流。国内有关网球比赛影响因素方面的文献共两百篇左右，这些文献主要使用回

归分析、判别分析、因子分析等方法，研究主要分为三类：关于某个或某些优秀网球运动员比赛影响因素的研究、对某项特定的网球赛事中球员比赛影响因素的研究、多项不同的网球赛事中球员比赛影响因素的综合研究等。

（1）特定网球运动员比赛影响因素的研究

对于特定的球员比赛影响因素的研究是很有必要的。在与一个球员交手之前，需要对他的技术特点、打法类型、获胜因素进行全面的了解。而且诸如费德勒、纳达尔、德约科维奇这样常年占据世界前三的顶级球员，他们实力强大，技战术多变，是所有的球员及相关专家重点研究的对象。

郭立亚、袁毅等人（2010）以 2008 年男子单打排名前一百位的选手的十项技术指标和排名积分为研究对象，通过偏相关分析和逐步回归分析建立了排名积分与技术指标的回归方程，最终得出影响男子单打比赛胜负的三个最关键的技术指标分别是：二发得分率、接发球胜率和挽救破发点成功率。球员应该通过科学专业的训练提高这三方面的能力[10]。祁航（2014）以纳达尔为主要研究对象，通过他和老对手费德勒、德约科维奇、穆雷交手的 89 场单打比赛数据，运用逐步判别分析法对数据进行处理，得出变量对比赛结果影响力强弱排序先后为：接二发回球得分率、二发得分率、一发得分率、接一发回球得分率、挽救破发点成功率[11]。

何文盛、张力为、张连成（2011）采集了 2001 年到 2008 年纳达尔、费德勒、德约科维奇三人与排名前 10 名的运动员共交手的 230 场比赛的技术统计数据，建立了世界前 3 名男子网球运动员比赛胜负预测的 Bayes 判别方程。该论文的最大不同之处就在于它分析了不同的场地类型对网球比赛的影响，这是国内其他网球论文进行影响因素分析时从来没有做过的。在世界排名前 3 的男子网球运动员中，一发得分率和接一发得分率是获胜的关键因素。不同的场地往往会制约某一特定运动员的制胜能力。作者全文对数据使用的分析方法非常严谨，值得我们去学习。但论文的研究对象只针对世界排名前三的顶级选手，得出的结论代表性并不强[12]。蒋启飞、郑贺（2015）以 2014.01.01 到 2014.10.25 期间排名前 43 位的男子网球单打选手的技术指标为研究对象。首先，对所收集整理的指标数据进行因子分析，计算每位选手在此期间因子得分；其次，根据所有选手获得的因子综合得分作为因变量进行技术回归预测模型的构建。通过本研究所建立的回归预测模型方程，准确率高达 90% ，所构建的回归预测模型方程具有较高的准确性、可信性，因此

用此模型对男子网球单打选手进行综合实力预测也是可行的[13]。

（2）特定赛事影响因素研究

不同于其他的球类运动，网球比赛按照不同的场地类型可分为硬地赛事、红土赛事、草地赛事和地毯赛事。场地的差异会导致球员发球时速、击球落点、回球高度的差异。不同类型的场地对球员的接发球、技战术水平、体能等有不同的要求，所以针对特定赛事影响因素进行研究可以帮助球员针对性的提升，制定更加恰当的比赛策略，具有较强的现实意义。

张银满（2009）选择 2008 年美网公开赛和 2009 年澳大利亚网球公开赛中的 132 场男子单打比赛作为研究对象，文章研究所使用的比赛统计指标是参照网球比赛转播时官方常用的技术统计指标，对有关的网球专家、教练及球员进行问卷调查，最终筛选确定的。运用单因素方差分析和逐步判别分析对数据进行处理，得到在男子网球比赛中对比赛胜负影响最大的因子依次为一发成功率、一发得分率、二发得分率以及接发球得分率[14]。

杨志敏（2010）的研究对象为他从 ESPN 与中央五台转播的 2008 年美网公开赛和 2009 年澳网公开赛中选取完整的 84 场男子单打比赛，他选取了胜负双方的 10 项技术指标，通过逐步判别分析，最终筛选得到了对比赛胜负影响最大的 4 个因素：一发成功率、一发得分率、二发得分率及接发球得分率。他还根据非标准化判别函数的判别结果，给出这四个影响因素的系数和常数，建立了关于比赛获胜的非标准化的判别方程。将 2008 北京奥运会男子单打比赛数据带入该判别方程，预测胜负的准确率在 94%以上[15]。罗伟权、张磊（2020）采用 Logistic 回归分析法，对 2019 年温布尔登网球公开赛男子单打比赛 254 场比赛构建基于 logistic 回归的职业网球运动员制胜因素判断模型。结果表明: 影响职业网球运动员制胜的关键因素为一发得分率、二发得分率、一发接发球得分率、二发接发球得分率和反手握拍类型。提出职业网球运动员要进一步提高发球得分率和接发球得分率，且使用双手反手握拍获胜的概率较高，这对于我国男子网球运动员的科学化训练具有一定的指导意义[16]。

（3）多项赛事影响因素研究

2012 年之后网球比赛影响因素的研究多集中于对多个球员参加的多项不同的赛事进行分析，这样的研究可以得出当今球员比赛时的技术特点，打法变化，以

及赢得比赛对球员技术的大致要求。

岳斌（2013）将 2008 年至 2012 年排名前 32 的选手划分为 4 个组别，第一组（排名 1-4）；第二组（排名 5-8）；第三组（排名 9-16）；第四组（排名 17-32）。根据 859 场硬地比赛的技术统计指标进行不同组别的对比分析，找出各组别获胜方和失败方之间的差异，并通过分析得出的相关因子用逐步回归的方式建立了胜负回归方程。他的方法与跟其他人相比起来多了一个组间比较，通过组间和组内的比较，并建立相关的回归方程得到接发球赢球率和二发赢球率是影响所各组(组内、组间)的最为关键的指标[17]。

在使用判别分析、回归分析等方法进行网球影响因素研究时，选择的初始技术指标很多都会因为共线性而被剔除，这导致很多文献的研究结果显示网球比赛只受到少数因素的影响，所以孟凡明、黄文敏（2019）选择了近 5 年四大满贯公开赛的 198 场女子单打比赛的 23 项技术统计数据，生成决策树指标重要性和决策树分支树。决策树分支树模型显示，41%的接发球得分率是女子网球比赛胜负的生命线，这些统计数据进一步说明了接发球得分率在现在女子网球比赛中的重要决定性。因此，加强接发球得分能力成了现在女子网球运动员的首要任务[18]。

1.3.2 国外研究现状

国外学者很早就开始进行网球相关的研究，他们对网球的研究涉及了球员的技战术水平分析、网球比赛的影响因素、网球赛果预测以及网球投注等各个方面，其中最主要集中于网球赛果的预测。并且国外十分重视各类网球比赛数据的收集，因此有数量庞大且指标详细的网球数据满足分析的要求，相关的文献也十分丰富，其中最为常见的就是以马尔可夫模型为代表的网球赛果预测模型。

Barnett, T.和 Brown, A.（2005）介绍了用 Markov 链基于每个球员发球得分的概率计算局、盘、比赛的获胜概率和预测比赛结果。他们在计算中假设球员发球得分的概率是独立同分布的随机变量，即在计算过程中发球得分概率是一个常数[19]。Newton（2005）考虑了 128 个球员参加比赛，采用单淘汰制的赛事，给出了每个球员获得冠军的概率（赢得每轮比赛的条件概率乘积），同时预测了半决赛四个球员获胜的概率[20]。Newton（2006）在计算时，球员的发球得分的概率每轮更新一次，每轮结束计算一次其击败对手的概率和获得冠军的概率[21]。Newton

（2009）在 2005 年的基础上，考虑了对手的接发球能力和整个赛季的接发球得分概率，校正了 A 与 B 对阵时，A 的发球得分的概率，然后将校正的概率用到前面的 Markov 链中[22]。Barnett、Clarke（2006）考虑了用修正的 Markov 链预测网球比赛结果，他考虑的选手发球得分的概率不是一个常数（文中用 2003 年澳大利亚公开赛的赛果预测说明发球得分概率是常数这一假定是不合适的），文中假设如果选手 A 领先，则 A 胜一盘的概率增加α，根据盘比分来调整盘得分的概率，但文中没有讨论α值如何选取[23]。

除马尔可夫模型外，现在国外的学者在这方面的研究越来越多的运用一些机器学习模型，如神经网络模型、逻辑回归模型等，时间序列模型和 Bradley-Terry 模型也被广泛用于网球赛果预测。

Amornchai S.，Suphakant P. 和 Chidchanok L.（2009）用神经网络预测网球比赛，文中用了三种多层模型：StatEnv 模型、AdvancedStatEnv 模型和 Timeseries 模型和反向传播算法。文中用 Barnett、Clarke（2006）的模型和 StatEnv 模型分别对 Australian2003 的比赛结果进行了预测，比较结果为 Barnett、Clarke（2006）的为 72.4%，StatEnv 模型 75.5906%。文中对 StatEnv 模型、AdvancedStatEnv 模型的预测结果比较表明：在训练数据足够多的时候，AdvancedStatEnv 模型的预测结果要比 StatEnv 模型的预测结果好。Timeseries 模型预测结果比 StatEnv 模型、AdvancedStatEnv 模型的结果要好，而且其结果表明选手在过去一年的表现对预测结果影响很大[24]。

Glickman（1999）将 Bradley-Terry 模型应用到网球比赛预测中，他给出了 1995 年赛季的男子网球排名，与 ATP 的官方排名相似，但是预测能力如何 Glickman 没有进行分析[25]。Ian Mchale 和 Alex Mortan(2011)中也将 Bradley-Terry 模型用到网球比赛预测中，他不仅考虑了胜的选手，而且还考虑了每个选手的胜的局数，文中同时还分析了场地的影响，每个选手在不同的场地有不同的表现。最后文中比较了 Bradley-Terry 模型和两个 Logit 模型：官方排名和两个选手的官方排名点（rank point）[26]。

Michal Sipko(2015) 从原始历史数据中提取 22 个特征，使用逻辑回归和人工神经网络两种机器学习算法的模型对网球比赛进行预测，并且使用神经网络模型进行网球投注，得到了 4.35%的投资回报[27]。Yixiong C(2018)的研究表明，在法网

这样的红土场地，接发球能力越强的球员表现越好。而在澳网和美网的这样的硬地场地，发球越好而且非受迫性失误越少的球员表现就越好[28]。

综上所述，国外关于网球运动的研究相较于国内比较成熟。国外的研究以很多选手为研究对象，可以通过研究总结出一些网球方面的普适性规律。而国内的网球研究大多针对具有代表性的运动员个体，对个案进行分析，集中于对网球比赛制胜因素方面的分析。同时国内的文献关于网球的研究，研究数据量小，研究方法方面比较单一，比较落后，因此得到的研究结果也有很大的局限，不具备很强的说服力。

在当今这样一个大数据的时代，各级各类的科学研究都注重大数据的概念，为此网球方面的研究更应该拓宽数据来源，不应仅仅局限于某一个球员的一些比赛或者是某站赛事，而是应该运用当下比较创新的方法，从大数据中得出网球运动的规律。

1.4 研究内容

本论文分为六个章节，各章节的主要内容如下：

第一章，绪论部分。主要介绍了本论文的研究背景及研究目的。通过对国内外学者关于网球影响因素和赛果预测文献的研究和归纳，引出了本文的研究内容和创新点。

第二章，相关理论与研究方法。简单介绍了网球比赛规则及网球数据的获取，详细介绍了本文所涉及的机器学习模型包括 logistic 回归、多层感知机模型、支持向量机模型，最后介绍了网球投注方面的理论知识。

第三章，网球比赛制胜因素分析及球员评价。以 2021 年 12 月 27 日 ATP 官方网站给出的男子网球单打排名前 100 的选手为样本，收集一百位运动员从参加职业赛事开始至 2021 年 12 月 27 日为止，期间所有的职业单打比赛指标数据作为样本数据。使用因子分析法找出男子网球比赛的制胜因素并根据球员得分进行球员评价。

第四章，基于排名系统的马尔可夫模型赛果预测。使用马尔可夫模型根据选手在各自发球局赢得一分的概率，递归地计算出选手赢得局、决胜局、盘和比赛胜利的概率，并介绍了常用的排名系统。提出改进的排名系统，使用该排名系统

来估计球员在发球时赢得一分的概率，并将这些概率输入到马尔可夫模型中，以获得球员赢得比赛的概率，并对排名系统及模型进行优化。

第五章，基于机器学习模型的网球赛果预测及网球投注。基于 2000 年到 2019 年的 ATP 比赛的所有数据，使用机器学习中的支持向量机模型、逻辑回归模型、以及多层感知机模型对网球赛果进行预测。比较各模型的性能，以确定最佳预测模型。开发一个简单的网球比赛投注模型。该模型使用 xgboost 模型和赔率数据对所选比赛进行投注，以获得最高的投资回报率。

第六章，总结与展望。对本文的研究工作进行充分的总结，并对未来的研究方向做出展望。

1.5 研究创新点

（1）创建图形工具。本文创建了一个图形工具，该图形工具可以随着时间的推移呈现球员的状态，帮助更好的了解球员的状态起伏和职业生涯的发展。它还可以指导职业球员的训练，比如如果图形工具显示球员的发球排名在提高，而接发球排名有下滑的趋势，球员本人以及教练就可以想办法在训练中通过增加接发球相关的各种训练来针对性的进行提高。不仅如此，在图形工具的帮助下，可以分析初始马尔可夫模型以寻找优化空间，并为应用和未来的工作提供了指导。

（2）网球建模方法。使用马尔可夫模型对网球赛事建模最重要的就是估计算球员在自己的发球局时赢得一分的概率。以往对球员发球得分概率的估计，大多都是将球员面对所有对手的统计数据进行平均，这种模型的表现非常糟糕，它存在较大的误差。本文创新性的同时利用马尔可夫模型与排名系统，形成了一种新的网球建模方法。使用改进后的排名系统，估计球员在发球局赢得一分的概率，然后将这些值输入马尔可夫模型计算球员赢得一局、一盘、以及一场比赛的概率。模型还考虑了网球比赛不同类型的场地影响以及发球接发球环节的内在差异，并将其纳入模型，对排名系统及模型作出改进，提高比赛预测的正确率。

（3）模型的应用。由于每场网球比赛的比赛时长都不是固定的，每站巡回赛必须估计媒体日程。根据这方面的需求，本文提出了一种使用蒙特卡罗模拟和正态分布拟合来估计网球比赛持续时间的简单方法。通过使用蒙特卡洛模拟，可以找到得分的分布，使用比赛时长和得分的线性方程将其转换为比赛持续时间，然

后拟合正态分布，因此得到的正态分布可用于得到比赛的预期长度。对于球迷关注的网球投注问题，本文则开发了一个简单的网球比赛投注模型。该模型使用xgboost 模型和赔率数据对所选比赛进行投注，以获得最高的投资回报率。

第二章 相关理论及研究方法

2.1 网球比赛

网球是一项比赛双方隔着球网、使用网球拍击打橡胶制空心球的球类运动，它起源于英国，网球比赛可以分为单打和双打，双打又有男双、女双、混双三种不同的类型。为了简单起见，本文的研究只专注于对男子网球的单打比赛进行影响因素分析与赛果预测。

网球比赛的任意一个回合，都是一名球员被指定为发球方，另一名球员被指定为接发球方，球员往往在发球时的表现更具优势。网球场是一个长方形的区域，场地上用白线划出分界线。不同于其他球类运动，网球比赛分别有红土、草地、硬地和地毯四种不同的场地类型。在发球方成功发球(有两次尝试机会)之后，双方轮流击球，直到一方赢得这个回合并获得一分。网球规则的全部细节由国际网球联合会在网上发布。

男子单打、双打采用五盘三胜或三盘两胜制，女子单打、双打以及男女混合双打则采用三盘两胜制。比赛时胜第一球记 15 分，胜第二球记 30 分，胜第三球记 40 分，再胜一球则为胜一局；当双方各得 40 分时为平分，平分后必须净胜 2 球才算胜一局；先胜 6 局者为胜一盘；当各胜 5 局时，一方必须净胜两局才算胜一盘[29]。

每年有 11 个月都在举办各种不同级别的的职业网球巡回赛，由职业网球联合会（ATP）和国际女子网球协会（WTA）分别组织各自的男子和女子网球巡回赛。本文只专注于男子网球巡回赛结果的分析预测，这并不会使本文的模型失去一般性，因为男子比赛与女子比赛的主要区别是增加了五盘三胜的可能性，而这在女子职业网球中是不存在的。

2.2 网球比赛数据

网球比赛的历史数据在网上随处可见。ATP 官网提供每一场比赛交手双方球员的个人信息、每场比赛的各项技术统计数据以及球员之间的历史交手记录。类似 tennis-data.co.uk 这样的网站，直接提供详细的整理成表格文件的网球数据。如

需要更复杂、时间跨度更久、技术统计数据更全面更准确的数据集，也可以在网上购买。

2.3 机器学习模型

对于特定的比赛，输入向量可以包含比赛和球员的各种特征，输出值可以是比赛的结果。相关特征的选择是构建成功的机器学习算法的挑战之一。可以用两种方法来处理网球赛果预测问题：

作为一个回归问题，其中输出是实值的。输出可以直接表示比赛获胜概率，但历史比赛的真实获胜概率是未知的，这迫使我们使用离散值来训练示例标签(例如，1 表示赢了比赛，0 表示输了比赛)。作为一个二元分类问题，还可以尝试将比赛分为“赢”或“输”两类。一些分类算法，如逻辑回归（如第 2.3.1 节所述），也给出了属于某个类别的实例的确定性的一些度量，可以用作比赛获胜概率。下面介绍本文在网球赛果预测时用到的三种机器学习算法。

2.3.1 逻辑回归模型

Logistics Regression 被称为逻辑回归，即使它的名字中有回归，但是逻辑回归实际上是一种分类算法，逻辑函数的性质是算法的核心。Logistic 函数$\sigma(t)$定义如下式所示：

$$\sigma(t)=\frac{1}{1+e^{-t}} \tag{2.1}$$

如图 2.1 所示，逻辑函数将$-\infty$和$+\infty$之间的实值输入映射为 0 和 1 之间的值，从而可将其输出解释为概率。

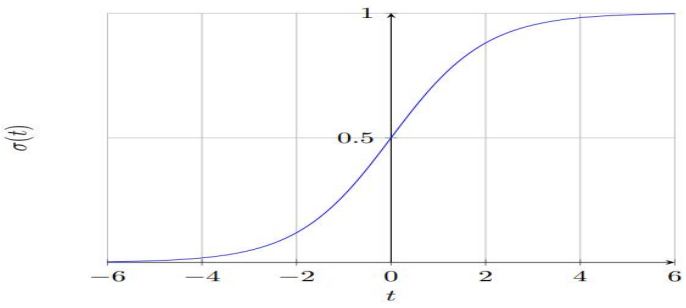

图 2.1 逻辑函数$\sigma(t)$

用于网球比赛预测的逻辑回归模型由 n 维比赛特征向量 $x=(x_1,x_2,\cdots,x_n)$ 和 n + 1 维实值向量 $\beta=(\beta_0,\beta_1,\cdots,\beta_n)$ 组成。为了使用模型进行预测，首先将 n 维特征空间中的一个点投影到一个实数上。

现在可以使用等式 2.2 中定义的逻辑函数将 z 映射为 0 到 1 内的概率值：

$$p=\sigma(z)=\frac{1}{1+e^{-z}} \tag{2.2}$$

该模型的训练包括优化参数 β，以使该模型为训练数据提供比赛结果的最佳再现。这是通过最小化 logistic 损失函数(方程 2.3)来实现的，该函数给出了模型在预测用于训练的比赛结果时的误差的度量。

$$L(p)=-\frac{1}{N}\sum_{i=1}^{N}p_i\log(y_i)+(1-p_i)\log(1-y_i) \tag{2.3}$$

其中：N 为训练比赛场次，p_i 为预测第 i 场比赛获胜的概率，y_i 为第 i 场比赛的实际结果（0 代表比赛输，1 代表比赛赢）。

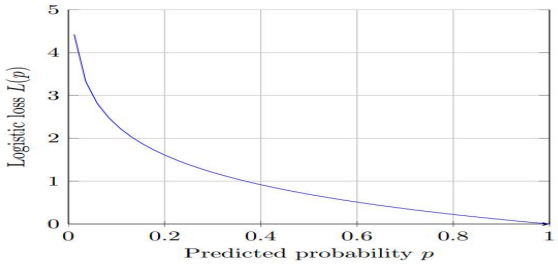

图 2.2 预测比赛获胜的逻辑损失

图 2.2 显示了假设比赛最终获胜，一场比赛基于不同的预测概率所导致的逻辑损失。任何偏离 $p=1.0$ 的正确预测的情况都将被惩罚。

2.3.2 多层感知机模型

多层感知机模型是由感知机模型推广而来的，也叫人工神经网络模型。它是一个由相互连接的“神经元”组成的系统，其灵感来自于生物神经元。人工神经网络每个神经元从其输入中计算出一个值，然后将其作为输入传递给其他神经元。它的结构通常有输入层、隐藏层与输出层。每个非输入层中的一个神经元连接到前一层的所有神经元。三层网络如图 2.3 所示。

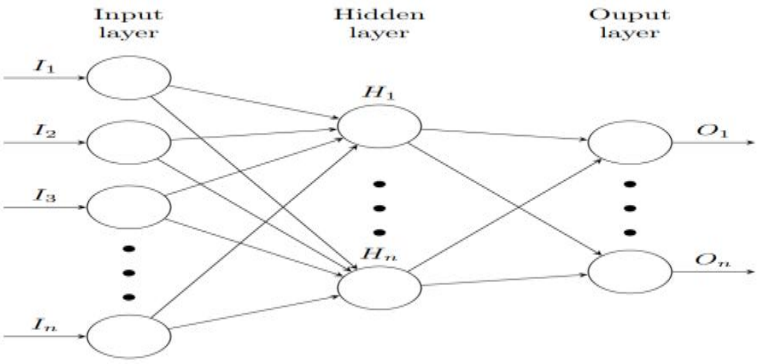

图 2.3 多层感知机模型的结构

多层感知机有一个非线性的激活层，更适合高维和线性不可分割的特征。其中每个神经元的结构都如图 2.4 所示，包含四个部分：

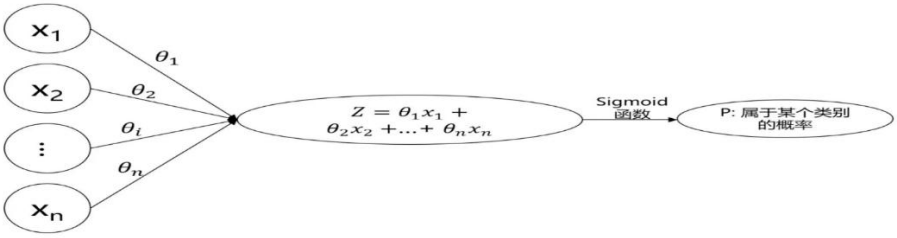

图 2.4 神经元的结构

（1） 输入：$x_1, x_2, \cdots, x_n$ 为神经元的输入，其表达式为：

$$X = [x_1, x_2, x_3, \cdots, x_n]^T \tag{2.4}$$

（2）加权求和：$\theta_1, \theta_2, \theta_3, \cdots, \theta_n$ 为神经元的权重，其表达式为：

$$\theta = [\theta_1, \theta_2, \theta_3, \cdots, \theta_n]^T \tag{2.5}$$

神经元节点首先对输入向量进行加权求和，得到隐式表征值：

$$z = \sum_{i=1}^{n} \theta_i x_i \tag{2.6}$$

（3）激活函数：用于对加权求和后的值进行函数运算，加强神经元节点的表征能力。其种类较多，例如 Sigmoid 函数、Tanh 函数及 Relu 函数[30]。

（4）输出：用于传递激活函数处理后的数值：

$$o = \sigma(z) = \sigma(\theta X) \tag{2.7}$$

式中，σ 代表激活函数。

本文使用的多层感知器是在 sklearn 中实现的，它有一个输入层，一个包含 100

个单元的隐藏层和一个输出层。选择 ReLU 作为激活函数，Adam 作为求解器。

2.3.3 支持向量机模型

与前面的机器学习模型一样，支持向量机模型也是监督学习模型。支持向量机是通过将样本映射为空间中的点，并找到一个最大间隔超平面，将它们划分为它们所标记的类别(如前所述，这些类别可以是“获胜”和“失败”类别)来构建的。

目前将支持向量机应用于网球比赛预测的研究还相对较少。在这种情况下，支持向量机相比多层感知机模型有几个优势。首先，训练不会产生局部最小值，这在多层感知机中很常见。此外，支持向量机在预测精度方面通常优于多层感知机模型，特别是当特征与训练样本的比率很高时。然而，支持向量机的训练时间往往要高得多，而且模型往往难以配置。

机器学习算法在实现的过程中，还常面临两个问题：

（1）过拟合。虽然有大量的历史数据可用于训练上述模型。然而，需要注意的是，球员在即将到来的比赛中的表现需要根据他们过去的比赛来估计。只有最近在同一场地与相似对手的比赛才能准确反映出球员的预期表现。因此，网球模型本身就存在数据不足的问题。数据的缺乏往往导致模型的过拟合，即模型描述的是数据中的随机误差或噪声，而不是潜在的关系。多层感知机模型特别容易过度拟合，特别是当隐层/神经元的数量相对于示例的数量较大时[31]。为了克服过拟合问题，只使用最相关的比赛特征进行训练。这些特征被选择的过程称为特征选择，有各种各样的算法。去掉不相关的特征也会提高训练时间。

（2）超参数优化。模型的训练优化了模型参数，如多层感知机模型中的权重。然而，模型通常也有超参数，这些超参数不是学习得到的，必须提供。例如，隐藏层的数量和每一层中的神经元数量是一个神经网络的一些可配置的超参数。对于给定的模型，得到最优超参数的过程往往是经验的。传统的算法，网格搜索，涉及到对预定义超参数空间进行彻底搜索。一个成功的网球预测模型需要仔细选择超参数。

2.4 网球投注

对体育博彩来说，网球赛事有一些有趣的特性。考虑到网球这种独特的得分

体系，在特定的一个比赛日，实力较弱的球员爆冷赢得比赛的情况是相当少见的。这说明网球是一项很容易预测的运动。然而，作为一项个人运动，不难想象，相较于团体运动来说，单个运动员的表现会比团体项目里一个团体的表现更不稳定，这反过来又使得网球比赛很难预测。这项运动的流行性和这种不可预测性使网球成为一个有吸引力的博彩市场。

由于在线博彩的普及，体育博彩的流行程度在过去几年里有了显著的增长。网球投注主要有两种类型：赛前投注和比赛进行中投注，区别在于赛前投注一旦比赛开始了就不能再投注。一场网球比赛甚至在开始前就有超过 20 种不同的投注选择，这并不罕见。每年世界各地都有大量的网球比赛，其计分规则也很适合进行投注。在比赛中，投注者可以对许多不同的选择下注：谁将赢得下一局/盘，比赛将进行多少局，谁会赢得比赛等等。然而，在本文中，唯一的重点将是在比赛开始之前预测比赛的获胜者。因为这种投注类型的赔率从历史上来看是最容易获得的，这使我们能够对本文的模型针对投注市场的表现进行更全面的评估。

市场上经营欧赔的博彩体系主要包括交易所和博彩公司这两大类。网球比赛的投注既可以通过博彩公司进行，也可以通过博彩交易所进行。交易所体系就是我们前面提到的以 Betfair 交易所为代表的一些通过柜台或网上交易买卖等形式而经营的博彩交易机构，代表的公司有：Betfair（英国）、WBX(伦敦)、BETDAQ(爱尔兰)、Redbet(英国)。博彩公司则包括几百家大大小小的公司，这些公司以立博、威廉希尔、韦德、BET365、澳门彩票等为代表，就是大家所说的“百家赔率”[32]。网球投注通常涉及到以下几个概念。

（1）赔率

在预测一场网球比赛的结果时，要达到 100%的准确率是不可能的。投注赔率被定义为给定玩家赢得比赛的概率(隐含概率)的倒数，代表投注者通过正确预测事件结果而获得的收益。显然，赔率越高，赢得比赛的概率就越低。庄家给每一个可能的赌注分配适当的赔率。这些赔率代表了一个赌注的不同结果的可能性。如果一个结果被认为不太可能发生，那么收益将会很高。相反，大众普遍预期的结果将会有更低的收益。赔率可以用十进制、分数式或正负值等形式表示。到目前为止，大部分欧洲国家所属的博彩公司采用“十进制”赔率，也叫小数赔率，英国博彩公司除外，他们使用分数赔率多一些。本文使用的赔率将始终以十进制

形式表示。以 2019 年温网决赛诺瓦克德约科维奇和罗杰费德勒之间的赔率为例。比赛开始前，德约科维奇获胜的赔率为 1.55，说明大家普遍认为德约科维奇赢得冠军的可能性更大。如果下注者正确预测了德约科维奇获得比赛的胜利，那么他们每押一百英镑就会得到一百五十五英镑(包括原本下注的一百英镑)。如果德约科维奇输了，下注者就会把一百英镑的赌注全部输给庄家。同样的推理也适用于罗杰费德勒，他获胜的赔率为 2.66。需要注意的是，可能的收益取决于下注时的赔率，因为庄家可能会随着时间的推移而改变他们的赔率。

（2）隐含概率

投注的赔率可以通过取赔率的倒数表示为概率。这个概率代表了根据庄家的预测结果发生的概率。在上面的例子中，德约科维奇和费德勒获胜的隐含概率分别为 64.5%和 37.6%。将投注赔率转换为隐含概率对投注者很有用，因为它有助于识别特定市场的价值。它指出了你需要在多久的时间内正确下注才能获得正的预期回报。因此，如果你知道一个结果的可能性比它的隐含概率更高，那么长期来看，押注这个结果总是有利可图的。

（3）博彩公司利润

上面例子中隐含概率的总和并不等于 100%。超过 100%的差额等于庄家的利润(在这个例子中是 2.1%)。这部分利润即为博彩公司的利润来源。

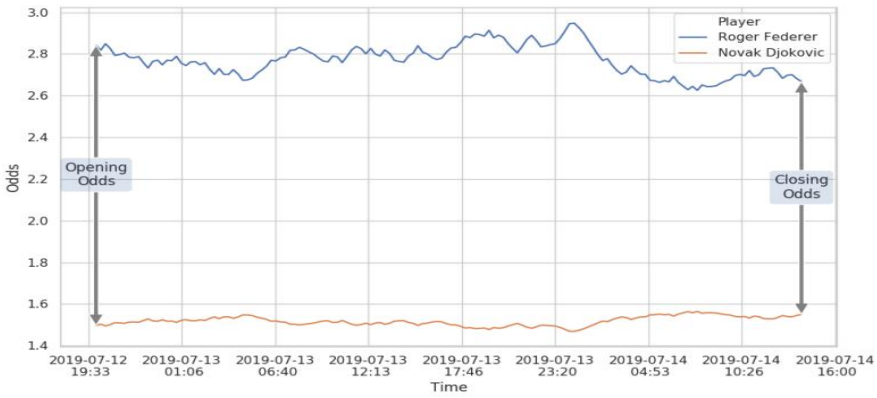

图 2.5 赔率调整

博彩公司决定每个球员的赔率有两个阶段。首先，他们根据大量的比赛数据对，甚至找人对赛前情况进行搜集，对自己认为的可能结果的真实赔率进行初步估计，并加上他们的利润。他们开出的赔率基本上反映了双方球迷的心理预期和下注情

况，不会有太大的偏差。这些赔率通常在网球比赛前一到两天公布，称为开局赔率。然后，随着下注资金的变化，博彩公司也会通过技术手段实时调整赔率，比如押球员 A 的人太多，系统就会调低球员 A 的赔率，提高球员 B 赢球的赔率，引导更多的资金去买球员 B 赢，再次形成平衡。无论比赛结果如何，他们都能获得利润。这种保证盈利的赔率和赌注被称为“荷兰赌”。这个设置赔率的动态过程如图 2.5 所示。

（4）博彩市场效率

终赔，是开赛前的临场赔率，是庄家在接收了大量投注，评估了各种动态因素后所最终确定下来的赔率。它们包含下注者所知道的所有信息和他们的意见。最终赔率更接近于网球比赛的真实赔率。最终赔率公布之后，可以获得关于这场比赛的更多信息，这些信息可能会影响预期的结果，比如：天气预报可能已经改变，特定的天气情况有利于某个球员，可能会出现关于某个球员受伤的消息等等。这些信息不一定需要公开。有内幕消息的投注者也可以影响赔率，只要他们下注的钱足够多。投注金额较大的庄家可以以较小的利润经营，因为他们的最终赔率更有效。图 2.8 以校准曲线的形式显示了Pinnacle的最终赔率的有效性。校准曲线使我们能够将模型的预测概率与事件的经验概率进行比较。在这种情况下，预测概率等于最终赔率的隐含概率。如果事件发生，经验概率等于 1，否则为 0。预测概率和经验概率用蓝线表示。

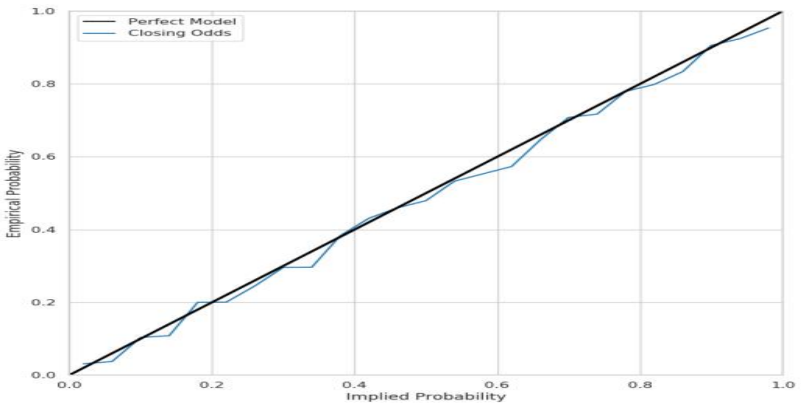

图 2.6 Pinnacle 的最终赔率校准图

从校准曲线可以清楚地看出，平均而言，隐含概率高于经验概率。正如第 2.4.2

节所解释的，只有当事件发生的真实概率高于概率的隐含概率时，下注才有利可图。因此，无论采用何种策略，在每一场比赛中投注相同的赌注都将导致负的预期回报。在考虑到数据窥探偏差的情况下，研究人员发现网球市场上 40 种不同博彩规则的表现并不盈利。

（5）击败博彩公司

显然，一个可获利的投注模型不应该对每一场比赛都下注。至少，一个好的模型应该在可信度高的情况下下更多的下注，在不确定的情况下下更少的下注。因此，如果目标是建立一个可以获利的投注模型，它不仅应该预测比赛的获胜者，而且还应该预测预测的可信度。总的来说，网球博彩市场非常高效，但这并不排除一些比赛存在误差。

第三章 网球比赛影响因素分析与球员评价

3.1 技术指标选取

该部分研究以2021年12月27日职业网球联合会官方网站给出的男子网球单打年度排名前100的选手为样本，收集这100位球员从开始参加职业网球比赛至2021年12月27日为止，整个职业生涯所有的单打比赛技术统计数据作为样本数据。

在网球比赛中，有许多的技术指标可能在不同的方面对于比赛的结果产生重要的影响。职业网球联合会官方网站关于每个球员的比赛数据中，共列出了十八个单打比赛的技术统计指标，这十八个指标按照不同的技术环节共分为以下三类，分别为：

发球环节：ACE球个数、双误个数、一发成功率、一发得分率、二发得分率、面临破发点个数、挽救破发点成功率、赢得发球局胜率以及发球局局数。

回球环节：接一发回球得分率、接二发回球得分率、破发机会、破发成功率、赢得接发球局胜率及接发球局的局数。

得分环节：发球得分率、接发球得分率以及总得分率。

初步分析，这十八个技术指标中，发球局局数和接发球局数对比赛的胜负并无较大影响，所以这两个技术指标不纳入比赛影响因素分析中，直接删除。剩下的十六个指标，分别从发球环节、接发球环节以及得分环节对比赛施加影响，这些技术指标内部彼此存在着一定的关系，他们互相影响，共同决定了比赛的走势。

再仔细观察，可以发现一发成功率、一发得分率、二发得分率、挽救破发点成功率、赢得发球局胜率、接一发回球得分率、接二发回球得分率、破发成功率、赢得接发球局胜率、发球得分率、接发球得分率以及总得分率这十四个指标均为比值型指标，考虑到指标之间的可比性及后续分析的方便性，应将剩余四个指标也相应的转化为比值型指标。转化的方式为：ACE球个数/发球局局数、双误个数/发球局数、面临破发点个数/发球局数、破发机会数/接发球局数。这样这四个指标就变成了发球局发出ACE球的概率、发球局发出双误的概率、发球局面临破发点的概率以及接发球局破发机会的概率。这四个转化后的指标与前面的十二个比值型指标选取合理，满足我们的分析要求，所以将这十六项技术指标作为网球比赛

因素分析的技术指标体系。具体的指标体系和变量解释如表 3.1。

表 3.1 变量说明表

	英文变量名	中文变量名	属性
发球环节	Aces/Service Games Played	发球局发出 ACE 球的概率	连续变量
	Double Faults/Service Games Played	发球局发出双误的概率	连续变量
	Break Points Faced/Service Games Played	发球局面临破发点的概率	连续变量
	1st Serve	一发成功率	连续变量
	1st Serve Points Won	一发得分率	连续变量
	2nd Serve Points Won	二发得分率	连续变量
	Break Points Saved	挽救破发点成功率	连续变量
	Service Games Won	赢得发球局胜率	连续变量
接发球环节	1st Serve Return Points Won	接一发回球得分率	连续变量
	2nd Serve Return Points Won	接二发回球得分率	连续变量
	Break Points Opportunities/Return Games Played	接发球局破发机会的概率	连续变量
	Break Points Converted	破发成功率	连续变量
	Return Games Won	赢得接发球局胜率	连续变量
得分环节	Total Service Points Won	发球得分率	连续变量
	Return Points Won	接发球得分率	连续变量
	Total Points Won	总得分率	连续变量

3.2 因子分析

在进行因子分析前，观察所选指标可以发现，发球局出现双误概率、发球局面临破发点概率这两个指标均为负向指标，取值都是越小越好，而其它指标都是取值越大越好，均为正向指标。为了保证所有指标在因子分析中对最终的总得分的贡献都是正向的，首先应该使用负数法对上述两个负向指标进行正向的转化，以保证所选的指标均为同趋势的。

首先进行第一次因子分析，根据分析的结果，只有接发球得分率指标的因子载荷小于 0.4，直接删除接发球得分率指标。再次进行因子分析，第二次因子分析

所有指标均满足条件，全部保留。最终确定了包括发球局发出 ACE 球的概率（Y_1）、发球局出现双误概率（Y_2）、发球局面临破发点概率（Y_3）、一发成功率（Y_4）、一发得分率（Y_5）、二发得分率（Y_6）、挽救破发点成功率（Y_7）、赢得发球局胜率（Y_8）、接一发回球得分率（Y_9）、接二发回球得分率（Y_{10}）、接发球局破发机会的概率（Y_{11}）、破发成功率（Y_{12}）、赢得接发球局胜率（Y_{13}）、发球得分率（Y_{14}）及总得分率（Y_{15}）在内的十五个指标作为网球比赛影响因素分析的技术指标体系。

对最终选定的十五项指标再进行一次因子分析。描述统计选择 KMO 与 Bartlett 球型检验，公因子的抽取方法选择主成分法，抽取因子的标准是特征值大于 1，旋转方法基于最大方差法，得分选择回归方法计算因子得分[33]。

（1）KMO 与 Bartlett 球型检验

如表 3.2 所示，KMO 检验值为 0.748，该值接近于 1，这表明所选十五项指标存在较强的的相关关系，而且偏相关性较弱，是非常适合进行因子分析的。Bartlett 球型检验结果 P=0.000，小于 0.01，拒绝原假设，说明相关系数矩阵不是单位阵，拒绝各变量独立的假设，两个检验结果均表明表明所选取的十五项指标适合做因子分析。

表 3.2 KOM 与 Bartlett 球形检验表

KMO 检验		0.748
Bartlett 球形检验	近似卡方	2791.916
	自由度	105
	显著性	0.000

（2）提取原始变量

如表 3.3 所示，发球局发出 ACE 球的概率（Y_1）、发球局面临破发点概率（Y_3）、一发得分率（Y_5）、赢得发球局胜率（Y_8）、接一发回球得分率（Y_9）、接二发回球得分率（Y_{10}）、接发球局破发机会的概率（Y_{11}）、赢得接发球局胜率（Y_{13}）、发球得分率（Y_{14}）及总得分率（Y_{15}）被提取的初始信息全部大于 0.8，这说明它们都可以被公因子非常好的进行表达。发球局出现双误概率（Y_2）、二发得分率

（Y_6）被提取的原始信息分别为 0.754、0.779，也可以被公因子表达得很不错。挽救破发点成功率（Y_7）、破发成功率（Y_{12}）、一发成功率（Y_4）被提取的原始信息介于 0.6 到 0.7，虽然被公因子提取的初始信息不够完整，有一部分的损失，但是均大于 0.5，说明都可以被表达。总的来说，这十五个指标被提取的信息较为完整，且基本都可以被公因子很好的表达，满足因子分析的要求。

表 3.3 公因子方差表

变量	起始	提取
Aces/Service Games Played	1.000	0.873
Break Points Opportunities/Return Games Played	1.000	0.910
1st Serve Points Won	1.000	0.970
2nd Serve Points Won	1.000	0.779
Break Points Saved	1.000	0.669
Service Games Won	1.000	0.983
Total Service Points Won	1.000	0.984
1st Serve Return Points Won	1.000	0.838
2nd Serve Return Points Won	1.000	0.842
Break Points Converted	1.000	0.605
Return Games Won	1.000	0.978
Double Faults/Service Games Played	1.000	0.754
Break Points Faced/Service Games Played	1.000	0.907
1st Serve	1.000	0.634
Total Points Won	1.000	0.962

（3）可解释方差的比重

如表 3.4 所示，所选取的十五个指标，前三个因子的特征值大于 1，且累计方差贡献率达到了 84.589%，所以提取前三个因子作为主因子，分别记为F_1、F_2、F_3。F_1的方差贡献率为 48.364%，F_2的方差贡献率为 25.120%，F_3的方差贡献率 11.104%。由图 3.1 的碎石图可以看出，前三个因子所处的斜率都是非常陡峭的，但是从第四个因子开始，斜率逐渐开始变缓，因此，此处选择前三个因子作为主因子是非常合理的。

表 3.4 方差贡献率表

	初始特征值			提取载荷平方和			旋转载荷平方和		
成分	总计	方差贡献百分比	累计方差贡献百分比	总计	方差贡献百分比	累计方差贡献百分比	总计	方差贡献百分比	累计方差贡献百分比
1	7.255	48.364	48.364	7.255	48.364	48.364	5.741	38.272	38.272
2	3.768	25.120	73.485	3.768	25.120	73.485	5.118	34.122	72.393
3	1.666	11.104	84.589	1.666	11.104	84.589	1.829	12.195	84.589
4	0.706	4.706	89.294						
5	0.492	3.277	92.572						
6	0.458	3.053	95.625						
7	0.285	1.897	97.521						
8	0.162	1.081	98.602						
9	0.107	0.716	99.318						
10	0.041	0.271	99.590						
11	0.028	0.188	99.777						
12	0.015	0.103	99.880						
13	0.010	0.067	99.947						
14	0.005	0.033	99.979						
15	0.003	0.021	100.000						

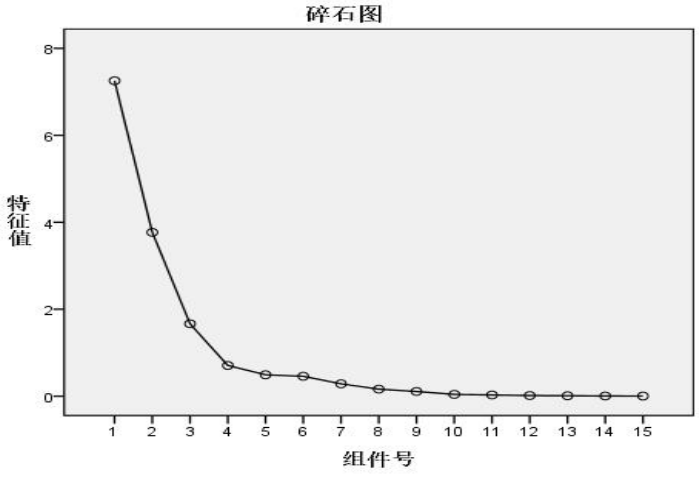

图 3.1 碎石图

（4）公因子载荷系数

如表 3.5 所示，F_1 在发球局发出 ACE 球的概率（Y_1）、发球局面临破发点概率（Y_3）、一发得分率（Y_5）、二发得分率（Y_6）、挽救破发点成功率（Y_7）、赢得发球局胜率（Y_8）、接一发回球得分率（Y_9）、接二发回球得分率（Y_{10}）、接发球局破发机会的概率（Y_{11}）、破发成功率（Y_{12}）、赢得接发球局胜率（Y_{13}）、发球得分率（Y_{14}）这些指标上载荷量比较高。F_2 发球局面临破发点概率（Y_3）、二发得分率（Y_6）、赢得发球局胜率（Y_8）、接一发回球得分率（Y_9）、接二发回球得分率（Y_{10}）、接发球局破发机会的概率（Y_{11}）、破发成功率（Y_{12}）、赢得接发球局胜率（Y_{13}）、发球得分率（Y_{14}）及总得分率（Y_{15}）这些指标上载荷量较高。F_3 在发球局出现双误概率（Y_2）、一发成功率（Y_4）、一发得分率（Y_5）、二发得分率（Y_6）方面载荷量较高。

表 3.5 成分矩阵

	因子		
	1	2	3
Aces/Service Games Played	0.903	-0.100	-0.220
Break Points Opportunities/Return Games Played	-0.731	0.596	-0.138
1st Serve Points Won	0.898	0.202	-0.351
2nd Serve Points Won	0.446	0.686	0.330
Break Points Saved	0.788	0.211	0.057
Service Games Won	0.893	0.429	-0.035
Total Service Points Won	0.891	0.437	-0.008
1st Serve Return Points Won	-0.728	0.508	-0.223
2nd Serve Return Points Won	-0.674	0.614	-0.105
Break Points Converted	-0.598	0.464	-0.179
Return Games Won	-0.742	0.640	-0.137
Double Faults/Service Games Played	-0.009	0.339	0.799
Break Points Faced/Service Games Played	0.833	0.458	-0.061
1st Serve	-0.115	0.170	0.769
Total Points Won	0.300	0.926	-0.123

分析可知，所提取的三个公因子的代表性指标并不显著突出，公因子的意义

并不明确，不能够很好的解释，因此需要通过最大方差法对所选的三个公因子进行旋转，以便更好的解释所提取的因子。

（5）因子旋转

表 3.6 旋转后的成分矩阵

	因子		
	1	2	3
Aces/Service Games Played	0.646	-0.606	-0.298
Break Points Opportunities/Return Games Played	-0.155	0.939	-0.069
1st Serve Points Won	0.851	-0.353	-0.347
2nd Serve Points Won	0.739	0.128	0.465
Break Points Saved	0.726	-0.372	0.056
Service Games Won	0.956	-0.261	0.016
Total Service Points Won	0.956	-0.261	0.045
1st Serve Return Points Won	-0.199	0.893	-0.036
2nd Serve Return Points Won	-0.104	0.906	-0.102
Break Points Converted	-0.134	0.766	-0.014
Return Games Won	-0.135	0.976	0.082
Double Faults/Service Games Played	0.117	0.061	0.858
Break Points Faced/Service Games Played	0.932	-0.195	0.003
1st Serve	-0.068	0.014	0.793
Total Points Won	0.836	0.926	0.098

如表 3.6 所示，旋转后的$F_{1转}$在发球局发出 ACE 球的概率（Y_1）、发球局面临破发点概率（Y_3）、一发得分率（Y_5）、二发得分率（Y_6）、挽救破发点成功率（Y_7）、赢得发球局胜率（Y_8）、发球得分率（Y_{14}）及总得分率（Y_{15}）这些指标方面具有较高的载荷，这些指标都是用于衡量发球局表现的指标，因此将其命名为发球局表现因子。旋转后的$F_{2转}$在接一发回球得分率（Y_9）、接二发回球得分率（Y_{10}）、接发球局破发机会的概率（Y_{11}）、破发成功率（Y_{12}）、赢得接发球

局胜率（Y_{13}）这些指标方面具有较高的载荷，这些指标都是用于衡量球员接发球局表现的，因此将其命名为接发球局表现因子。$F_{3转}$在发球局出现双误概率（Y_2）、一发成功率（Y_4）两个指标上具有较高的载荷，这两个指标可用于衡量球员发球的稳定性，因此将其命名为发球稳定性因子。而$F_{1转}$、$F_{2转}$、$F_{3转}$分别代表公因子F_1、F_2、F_3使用最大方差法进行旋转后得到的旋转后的公因子。

（6）因子得分

表 3.7 成分得分系数矩阵

	因子		
	1	2	3
Aces	0.093	-0.068	-0.142
Break Points Opportunities	0.035	0.200	-0.032
1st Serve Points Won	0.153	0.008	-0.198
2nd Serve Points Won	0.140	0.045	0.234
Break Points Saved	0.114	-0.038	0.040
Service Games Won	0.169	0.007	0.000
Total Service Points Won	0.168	0.005	0.017
1st Serve Return Points Won	0.026	0.195	-0.088
2nd Serve Return Points Won	0.042	0.193	-0.013
Break Points Converted	0.029	0.168	-0.067
Return Games Won	0.041	0.209	-0.029
Double Faults	0.000	-0.048	0.486
Break Points Faced	0.169	0.022	-0.012
1st Serve	-0.037	-0.067	0.458
Total Points Won	0.198	0.168	-0.011

根据表 3.7 得出的因子得分系数，计算网球比赛影响因素总得分，计算公式如下：

$$\begin{aligned} F_{1总得分} &= 0.093Y_1 + 0.035Y_2 + 0.153Y_3 + \cdots\cdots + 0.198Y_{15} \\ F_{2总得分} &= -0.068Y_1 + 0.200Y_2 + 0.008Y_3 + \cdots\cdots + 0.168Y_{15} \\ F_{3总得分} &= -0.142Y_1 - 0.032Y_2 - 0.198Y_3 - \cdots\cdots - 0.011Y_{15} \\ F_{总得分} &= \left(0.38272F_{1总得分} + 0.34122F_{2总得分} + 0.12195F_{3总得分}\right) / 0.84589 \end{aligned} \tag{3.1}$$

其中$F_{1总得分}$代表发球局表现因子得分，$F_{2总得分}$代表及接发球局表现因子得分，$F_{3总得分}$代表发球稳定性因子得分，$F_{总得分}$代表网球比赛影响因素总得分。在男子单打比赛影响因子中，发球局表现因子占38.272%，接发球局表现因子占34.121%，发球稳定性因子占12.195%。

3.3 球员分析

本章选择一百名网球运动员的数据进行影响因素的研究，主要目的是使得研究的结果更具有普适性，但是这一百名网球运动员的影响因素总得分数据非常的庞杂，所以只选取其中总得分排名前十六名的运动员进行分析。之所以选择前十六名进行分析是因为这些球员代表着世界网球的最高水平，而且这里只给出因子总得分排名前十六的球员并不会对后续研究结果产生任何不好的影响，同时也可以满足该部分的研究目的。

表3.8 世界前16名男子网球单打选手得分能力表

排名	姓名	$F_{1总得分}$	$F_{2总得分}$	$F_{3总得分}$	$F_{总得分}$
1	Rafael Nadal	2.19295	2.80407	1.62763	2.36
2	Novak Djokovic	2.02777	2.49856	0.57026	2.01
3	Roger Federer	2.61244	1.44153	0.60361	1.85
4	Carlos Alcaraz	0.29250	1.95928	0.45311	0.99
5	Daniil Medvedev	1.27855	1.25994	-0.9564	0.95
6	Bautista Agut	0.59641	1.19255	1.29811	0.94
7	Jannik Sinner	0.79101	1.34898	0.22653	0.93
8	Kei Nishikori	0.57511	1.39413	0.13536	0.84
9	Stefanos Tsitsipas	1.57188	0.09968	0.58951	0.84
10	Richard Gasquet	0.84768	0.79760	0.18289	0.73
11	Milos Raonic	2.54284	-1.0915	-0.1745	0.69
12	Alexander Zverev	0.92052	0.91388	-0.9443	0.65
13	Casper Ruud	0.68943	0.47222	0.95750	0.64
14	Marin Cilic	1.30158	0.45585	-1.0202	0.63
15	Brandon Nakashima	1.25184	-0.1242	0.54808	0.60
16	Dominic Thiem	0.90607	0.49582	-0.1922	0.58

如表 3.8 所示，通过总得分可以将得分排名前十六的球员划分为三档。总得分在 1.8 以上的球员为实力排名第一档，他们分别是 Rafael Nadal（纳达尔）、Novak Djokovic（德约科维奇）、Roger Federer（费德勒），这三位球员毫无疑问是当今男子网坛乃至男子网球历史上最伟大的三位球员，他们的得分是断层式领先于其他人的。实力排名第二档的球员总得分都是在 0.7-1 之间，依次分别是 Carlos Alcaraz（阿尔卡拉斯）、Daniil Medvedev（梅德韦杰夫）、Roberto Bautista Agut（阿古特）、Jannik Sinner（伊斯内尔）、Kei Nishikori（锦织圭）、Stefanos Tsitsipas（西西帕斯）及 Richard Gasquet（加斯奎特）。Milos Raonic（拉奥尼奇）、Alexander Zverev（兹维列夫）、Casper Ruud（鲁德）、Marin Cilic（西里奇）、Brandon Nakashima（中岛布兰登）及 Dominic Thiem（蒂姆）为实力排名第三档的球员，他们的得分都在 0.5-0.7 之间。这 16 名球员虽然代表着男子网球的最高水平，但他们每个人在发球局表现、接发球局表现、发球稳定性方面擅长的领域却各不相同。

在实力排名第一档的球员中：纳达尔是三巨头中唯一一个发球稳定性因子得分超过 1.5 的人，此外他的发球局表现因子得分、接发球局表现因子得分也高达 2.19295 和 2.80407，这说明纳达尔在三个方面均表现强势，他的技术非常的全面均衡，没有弱环，他的总得分为 2.36，在一百名选手中排名第一。德约科维奇拥有和纳达尔一样强势的发球局和接发球局表现，他的发球局表现因子得分和接发球局表现因子得分分别为 2.02777 和 2.49586，但是发球稳定性这一项相比于纳达尔他处于劣势，得分仅为 0.57026，有待于提高。最终他以 2.01 的总分排名第二。费德勒的发球局表现因子得分、接发球局表现因子得分、发球稳定性因子得分分别为 2.61244、1.44153、0.60361，这说明费德勒拥有他们三个人当中最棒的发球局表现，但是他的接发球发现和发球稳定性方面均为劣势，最终他得分 1.85 排名第三。

在实力排名第二档的球员中：阿尔卡拉斯作为一名 00 后的年轻小将位列第二档球员的第一名，他的三项得分分别为 0.29250、1.95928、0.45311，其中他的接发球表现因子得分甚至高于费德勒，这表明他有很强的打接发球局的能力，但是他的弱点也很明显，那就是发球局表现和发球稳定性都较弱，有待提高。排名第二档第二名的梅德韦杰夫的三项得分分别为 1.27855、1.25994、-0.95644，说明梅德韦杰夫虽然打发球局和接发球局的能力都很强，但是发球稳定性很差，他这一

项得分是前十名选手里唯一一个得分为负的，需要重点提高。阿古特三项得分分别为 0.59641、1.19255、1.29811，说明阿古特发球稳定性很棒，打接发球局的能力也比较强，但是打发球局能力比较弱，他最终得分为 0.94，在实力第二档球员中排名第三。其余球员可以类似分析。

实力排名第三档的球员中：排名第一的拉奥尼奇有着堪比第一档球员的打发球局的能力，但是打接发球局的能力以及发球稳定性都较差。排名第二的兹维列夫打发球局和接发球局的能力都处于中等，但是发球稳定性很差。鲁德在第三档球员中最为全面，各方面得分都很平均。剩余球员类似分析。总的来说，实力处于第二档和第三档的球员，有些是有明显的技术缺陷，有些虽然各方面表现相当，但水平相对不高，相较于前三个最伟大的球员，他们需要提高自己的全面性，才有可能追赶得上前辈。

第四章 基于排名系统的马尔可夫模型赛果预测

4.1 马尔可夫链概率计算

根据 2.1 节的网球计分规则，网球比赛分为分、局及盘。一局比赛由分构成，一盘比赛由不同的局构成，而一场比赛由不同的盘构成。这种结构非常适合用马尔可夫法进行建模。一旦有了两名选手在各自发球局赢得一分的概率，就可以递归地计算出选手赢得局、抢七局、盘和比赛胜利的概率。需要注意的是，使用马尔可夫模型研究过程中，假设每一分都是独立同分布的。接下来用马尔可夫链基于每个球员发球得分的概率计算局、抢七局、盘、比赛的获胜概率。

（1）计算一局比赛的获胜概率

首先从一盘中的一局比赛来进行分析。在一局比赛中，如果一位球员首先得到四分或者四分以上，并且对对手多得两分，这位球员就可以获得一局比赛的胜利。一局比赛的马尔可夫模型如图 4.1 所示。

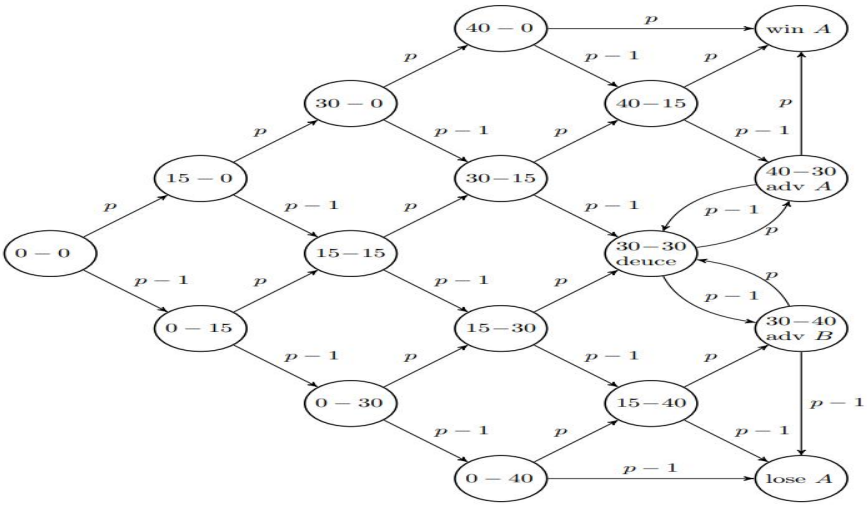

图 4.1 一局比赛马尔可夫模型

p 表示当球员 A 发球时，球员 A 赢 B 一分的概率，$1-p$ 表示球员 A 发球时，球员 B 赢 A 一分的概率。记当球员 A 发球时，他以 (a,b) 的比分赢球员 B 的概率为 $P(a,b)$。$P(a,b)$ 可以递归的写为：

$$P(a,b) = pP(a+1,b) + (1-p)P(a,b+1) \tag{4.1}$$

如果$a=4,b\le 2$，边界值是$P(a,b)=1$；如果$b=4,a\le 2$，边界值$P(a,b)=0$。当比分为$(3,3)$时出现问题，在这种情况下，我们用上述边界提出的问题无限递归。尽管如此，仍可以明确地计算$P(3,3)$的值。

$$P(3,3)=p^2P(5,3)+2p(1-p)P(4,4)+(1-p)^2P(3,5) \tag{4.2}$$

将$P(3,3)=P(4,4)$、$P(5,3)=1$以及$P(3,5)=0$代入得：

$$\begin{aligned} P(3,3)&=p^2+2p(1-p)P(3,3)\\ (1-2p(1-p))P(3,3)&=p^2\\ P(3,3)&=\frac{p^2}{p^2+(1-p)^2} \end{aligned} \tag{4.3}$$

表 4.1 球员 A 以不同的比分中赢球员 B 一局的概率

		B				
		0	15	30	40	game
A	0	0.62	0.46	0.27	0.10	0.00
	15	0.76	0.61	0.41	0.18	0.00
	30	0.88	0.77	0.60	0.33	0.00
	40	0.96	0.92	0.82	0.60	
	game	1.00	1.00	1.00		

表 4.1 显示了当$p=0.55$时所有可能的得分(a,b)的概率。

类似于p和P，q是当球员 B 发球时，B 赢得一分的概率，Q是球员 B 发球时赢得一局的概率。

（2）计算抢七局获胜的概率

在对一盘比赛建模之前，还有一种特殊的情况需要建模。当一局比分为 6-6 时，将进行抢七局的争夺。在进行抢七局比赛时，一名选手发球第一分，第二名选手发球后两分。之后，选手们每打两分交换一次发球。首先在抢七局中获得 7 分并且领先对手 2 分的球员将赢得抢七局和这盘比赛的胜利。

对抢七局的建模与普通局非常相似，见公式 4.1。为了区分普通局和抢七局的不同，当提到抢七局比赛，在概率上加上上标 T。使用与前面相同的表示法：

$$P^T(a,b)=pP^T(a+1,b)+(1-p)P^T(a,b+1),if\,(a+b)\div 2(\bmod 2)=0 \tag{4.4}$$

$$P^T(a,b)=qP^T(a,b+1)+(1-q)P^T(a+1,b),if(a+b)\div 2(\bmod 2)=1 \quad (4.5)$$

当$a=7,b\le 5$时，边界值为$P(a,b)=1$；当$b=7,a\le 5$时，边界值为$P(a,b)=0$。使用与方程 2.2 相同的逻辑，我们计算：

$$P(6,6)=\frac{p(1-q)}{p(1-q)+q(1-p)} \quad (4.6)$$

（3）计算一盘比赛的获胜概率

用P^S表示球员 A 赢一盘的概率。当球员 A 率先发球时，用接下来的两个方程来建模：

$$P^S(a,b)=P*P^S(a+1,b)+(1-P)*P^S(a,b+1),if(a+b)(\bmod 2)=0 \quad (4.7)$$

$$P^S(a,b)=Q*P^S(a,b+1)+(1-Q)*P^S(a+1,b),if(a+b)(\bmod 2)=1 \quad (4.8)$$

当$a=6,b\le 4$时，边界值为$P^S(a,b)=1$；当$b=6,a\le 4$时，边界值为$P^S(a,b)=0$。现在剩下未知的边界条件$P(5,5)$。当集合分数为 5-5 时，有几种可能的情况，但只有两种情况使球员 A 获胜。要么球员 A 连胜两局，要么球员 A 赢了一局，输了接下来的两局，然后赢了决胜局。可以用以下方程来表示：

$$P^S(5,5)=PQ+(PQ+(1-P)(1-Q))P^T \quad (4.9)$$

（4）计算一场比赛的获胜概率

有了球员 A 对球员 B 赢得一盘比赛的概率，对比赛进行建模就很简单了。将球员 A 获胜的概率表示为P^M，写出下一个方程：

$$P^M(a,b)=P^SP^M(a+1,b)+\left(1-P^S\right)P^M(a,b+1) \quad (4.10)$$

三局两胜的比赛，当$a=2,b\le 1$时，边界值为$P^M(a,b)=1$；当$b=2,a\le 1$时，边界值为$P^M(a,b)=0$。五局三胜的比赛，当$a=3,b\le 2$时，边界值为$P^M(a,b)=1$；当$b=3,a\le 2$时，边界值为$P^M(a,b)=0$。

（5）网球比赛方程

一旦获得了球员在自己的发球局时赢得一分的概率(p和q)，就可以使用如下方程来计算球员赢得一场比赛的概率。

获得一局比赛胜利的概率是：

$$P=\frac{p^4*\left(15-4*p-10*p^2\right)}{p^2+(1-p)^2} \quad (4.11)$$

获得决胜局胜利的概率是：

$$P^T = P^T(0,0) = \sum_{i=1}^{28} A(i,1)p^{A(i,2)}(1-p)^{A(i,3)}q^{A(i,4)}(1-q)^{A(i,5)}d(p,q)^{A(i,6)} \quad (4.12)$$

$$d(p,q) = \frac{p(1-q)}{p(1-q)+q(1-p)} \quad (4.13)$$

A 的值见附录 A.2。

获得一盘比赛胜利的概率是：

$$\begin{aligned} P^S = &\sum_{i=1}^{21} B(i,1)P^{B(i,2)}(1-P)^{B(i,3)}P^{B(i,4)}(1-P)^{B(i,5)} \\ &\times\left((PQ+(P(1-Q))+(1-P)Q)P^T\right)^{B(i,6)} \end{aligned} \quad (4.14)$$

B 的值见附录 A.1。

赢得一场比赛胜利的概率是：

$$\begin{aligned} P^M = P^M(0,0) &= \left(P^S\right)^2\left(1+2\left(1-P^S\right)\right), && \text{if best of 3} \\ P^M = P^M(0,0) &= \left(P^S\right)^3\left(1+3\left(1-P^S\right)+6\left(1-P^S\right)^2\right), && \text{if best of 5} \end{aligned} \quad (4.15)$$

如何估计球员在自己发球时赢得一分的概率(p 和 q)是马尔可夫模型进行赛果预测最大的难点。Barnet[19]的估计方法是将球员面对所有对手的统计数据进行平均，这样的估计存在较大的误差。因为实力强的球员通常更多的与实力强的球员进行对抗，而实力弱的球员更经常与实力弱的球员对抗。这意味着实力强的球员和实力弱的球员可能拥有非常相似的统计数据，但他们的真实技能水平却存在很大差异。

Knottenbelt 使用共同对手模型来估计输入，这样的模型可以较好地估计 p 和 q。在文章的最后，Knottenbelt 还讨论了比较所有球员之间交手的共同对手的可能性[34]。本文选择使用排名系统来估计 p 和 q 的值。

4.2 排名系统

在体育和其他竞技活动中，排名系统用于比较选手的实力。排名系统会根据选手的表现给他们相应的积分奖励。例如，根据赛事的级别和他们的排名，高尔夫球手会获得一定数量的积分。这些积分将在一个赛季结束前保留。现有许多不

同的排名系统，但本文的研究对采用数学方法的排名系统更感兴趣。本节主要介绍了 Elo 等级系统和 Glicko 排名系统，并对 Glicko 排名系统进行修正，提出改进后的 Glicko 排名系统。

4.2.1 Elo 等级系统

Elo 等级系统（Elo rating system）是指由 Arpad Elo 首先提出的一个用于对各类对弈活动水平进行衡量的评价方法，如今在对各类对弈水平进行评估时，它是学术界公认的比较权威的标准，已经被广泛应用于国际象棋、围棋、篮球、足球等运动中。现在深受广大年轻群体喜爱的各种网游比赛的匹配对战系统也采用 Elo 等级系统[35]。

Elo 等级系统是基于统计学的一个评估选手水平的方法。美国国际象棋协会最早开始使用这种系统。由于 Elo 等级系统相较以往使用的各种评价方法更加的公平客观，所以得到了广泛的认可与普及应用。Elo 模型原先采用的是正态分布。但是经过实践证明了，各选手的表现并不是呈现正态分布的，所以对此进行了修正，现在的等级系统通常使用的是逻辑分布。

Elo 排名对于数学模型很有帮助，因为它不仅能够对选手进行排名，还可得到一个选手战胜另一个选手的概率，使用前面图 2.3 所示的逻辑曲线估计该概率。假设选手 A 和选手 B 的当前等级分别为 R_A 和 R_B，则按照逻辑分布，选手 A 选手 B 的胜率期望值应当为：

$$E_A = \frac{1}{1+10^{(R_B-R_A)/400}} \tag{4.16}$$

类似选手 B 对选手 A 的胜率为：

$$E_B = \frac{1}{1+10^{(R_A-R_B)/400}} \tag{4.17}$$

选手 A 和选手 B 的比赛结束后，选手在比赛中的真实得分为 S_A，和他的胜率期望值 E_A 不同，则他的等级分要作相应的调整。具体的数学公式为：

$$R_A^{'} = R_A + K\left(S_A - E_A\right) \tag{4.18}$$

公式中 R_A 和 $R_A^{'}$ 分别为选手 A 调整前后的等级分。其中 K 为修正因子。关于 K

的取值，在国际象棋比赛中，通常取其值为 16，在其他大部分的比赛中，通常取值为 32。一般来说，对战水平越高的比赛中，K 的取值就越小，之所以这样做是为了避免少数的几场比赛就能改变顶级水平的选手的排名。这个参数取值应针对使用的特定模型进行优化。

现在假设 R_A 表示球员 A 的发球能力，R_B 表示球员 B 的接发球能力，球员 A 在发球局赢得一分的概率 p 用 E_A 来表示。这将是本文基于排名系统的马尔可夫模型的主要思想。

4.2.2 Glicko 排名系统

虽然 Elo 排名系统已经被实践证明是非常有效的，但 Elo 系统的问题在于其无法确定选手评分的可信度，而 Glicko 系统正是在此基础上进行改进的。假设两名评分均为 1700 的选手 A、B 在进行一场对战后 A 获得胜利，在美国国际象棋联赛的 Elo 评分系统下，A 选手评分将增长 16，对应地 B 选手评分将下降 16。但是假如 A 选手是已经很久没有参加比赛，但 B 选手每周都会参加比赛，那么在上述情况下 A 选手的 1700 评分并不能十分可信地用于评定其实力，而 B 选手的 1700 评分则更为可信[36]。

虽然很多情况下并不是这么极端，但是把选手评分的可信度考虑进入是很有必要的。因此 Glicko 系统扩展了 Elo，它不仅计算选手得评分（可以视为选手实力的“最佳猜测”），还加入了“评分误差”（RD，ratings deviation），从统计的角度来看，RD 是用于衡量一个选手得评分的不确定程度（RD 值越高，评分可信度越低）。如果一个选手的 RD 值很高，这意味着该选手并没有频繁地参加比赛，相反如果一个选手的 RD 值很低，则说明该选手参加比赛的次数很多。

在 Glicko 系统中，一个选手的评分之可能会因为他的比赛结果仅而发生变化，但其 RD 值的改变，不仅取决于他的比赛结果，还取决于他未参加比赛的时间长度。Glicko 系统的一个特点是，比赛的结果经常会使得选手的 RD 值减少，而未参加比赛的时间则经常会导致选手的 RD 值增加。关于这个现象的解释是：一个选手参加的比赛越多，就可以获取到越多关于该选手能力的信息，对他的评分也就越真实；而随着时间的流逝，对选手的真实实力就越不确定，这种不确定反映在 RD 值上就是增加。

正如刚才所提到的，在 Glicko 系统中，选手评分的变化并不像 Elo 那样经常是相同的。Elo 系统中，如果选手 A 的评分增长了 Y，则其对手 B 的评分会减少 Y，而在 Glicko 系统中并非如此。事实上，在 Glicko 系统中，对手 B 的评分究竟减少了多少取决于双方的 RD 值。

为了应用该算法，需要对发生在同一个“评分周期（rating period）”的所有比赛进行计算。一个评分周期可以长达数月，也可以短到一分钟。在评分周期的时间长度由相关人员自行设定。若选手没有评分，则其评分通常被设为 1500，评分标准差为 350。在下一小节，对 Glicko 排名系统作出改进以满足后续研究的需求。

4.2.3 改进的 Glicko 排名系统

正如上一小节所讨论的，Glicko 通过添加另一个变量，即球员评级的可靠性，升级了 Elo 排名系统。在 Glicko 排名系统中，每个球员被分配两个值，评级 r 和评级可靠性RD。RD取决于球员的活动。与参赛少不活跃的球员相比，经常参加比赛的球员具有更高的评分可靠性。Glicko 使用以下方程式计算 RD：

$$RD = \min\left(\sqrt{RD_{old}^2 + c^2 t}, 350\right) \tag{4.19}$$

其中t为自上次比赛至现在的时间长度（评分周期数），常数c是通过球员在一定时间范围内的表现的不确定性计算得到的，可以通过数据分析计算，也可以通过估算球员的评分标准差要多久能达到未评分球员的评分标准差而得到，350 则是新球员的评分标准差。

通过改变RD的计算方式来修改 Glicko 排名方法。为了计算参加过 m 场比赛的球员的RD，使用以下等式：

$$RD = \min\left(RD_{\text{begin}} - mRD_{\text{step}}, RD_{\text{end}}\right) \tag{4.20}$$

由图 4.2 可以看出，RD 以线性函数开始，到达RD_{end}时变成常数。图 4.2 展示了$RD_{begin} = 300$、$RD_{end} = 100$ 和$RD_{step} = 20$ 时，RD 与参加比赛数m的函数关系。

一旦计算出 RD，就可以在每场比赛的基础上修改排名。为了在与球员 2 的比赛后更新球员 1 的等级，使用等式 4.21 进行计算。

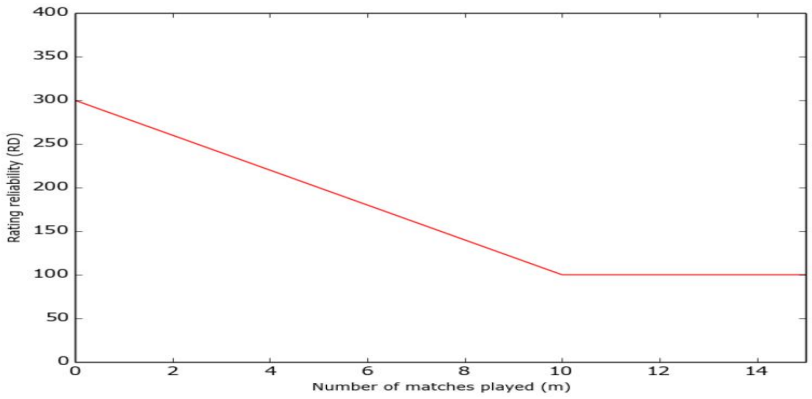

图 4.2 RD 与比赛次数 m 的函数

$$
\begin{aligned}
r_1' &= r_1 + \frac{q}{1/RD^2 + 1/d^2} g\left(RD_2\right)\left(s - E\left(s|\ r_1, r_2, RD_2\right)\right) \\
RD &= \min\left(RD_{\text{begin}} - mRD_{\text{step}}, RD_{\text{end}}\right) \\
q &= \frac{\ln 10}{400} = 0.0057565 \\
g(RD) &= \frac{1}{\sqrt{1 + 3q^2\left(RD^2\right)/\pi^2}} \\
E\left(s|\ r_1, r_2, RD_2\right) &= \frac{1}{1 + 10^{-g(RD_2)(r_1 - r_2)/400}} \\
d^2 &= \left(q^2 g\left(RD_2\right)^2 E\left(s|\ r_1, r_2, RD_2\right)\left(1 - E\left(s|\ r_1, r_2, RD_2\right)\right)\right)^{-1}
\end{aligned}
\tag{4.21}
$$

改进的 Glicko 排名系统还有另一个巧妙的优点：可以通过修改参数以适应不同的分析类型。如果想要进行球员职业生涯分析或一站锦标赛分析，则可以相应地修改 RD_{end}。较低的 RD_{end} 对于职业生涯分析更好，而较高的 RD_{end} 可以适合在较小的时间范围内分析变化。

下面给出一个使用改进的 Glicko 排名系统的示例。首先将参数设置为 $RD_{begin} = 300$、$RD_{end} = 100$ 和 $RD_{step} = 20$。

球员 1 参加了了 20 场比赛，等级为 1500。球员 2 参加了 6 场比赛，等级为 1700。接下来可以计算两位球员的评级可靠性：

$$
\begin{aligned}
RD_1 &= \min(300 - 20*20, 100) = 100 \\
RD_2 &= \min(300 - 6*20, 100) = 180
\end{aligned}
\tag{4.22}
$$

球员之间比赛，球员 1 对球员 2 得 0.75 分。现在更新球员的评分。正如预期，

球员 1 的等级评分将提高，而球员的等级评分则会降低。而且，球员 2 的评级将发生更大的变化，因为他的等级评分不太可靠。

计算球员 1 的评分变化：

$$
\begin{aligned}
g\left(RD_2\right) &= \frac{1}{\sqrt{1+3q^2\left(RD_2^2\right)/\pi^2}} \\
&= \frac{1}{\sqrt{1+3*0.0057565^2*180^2/\pi^2}} \\
&= 0.868302275 \\
E\left(s|\ r_1, r_2, RD_2\right) &= \frac{1}{1+10^{-g(RD_2)(r_1-r_2)/400}} \\
&= \frac{1}{1+10^{-0.8683*(-200)/400}} \\
&= 0.269006320 \\
d^2 &= \left(q^2 g\left(RD_2\right)^2 E\left(s|\ r_1, r_2, RD_2\right)\left(1-E\left(s|\ r_1, r_2, RD_2\right)\right)\right)^{-1} \\
&= \left(0.0057565^2*0.895786^2*0.269*(1-0.269)\right)**-1 \\
&= 203547 \\
r_1^{'} &= 1500+\frac{0.0057565}{1/100^2+1/203547}*0.8683*(0.75-0.269) \\
&= 1523
\end{aligned}
\tag{4.23}
$$

为了获得球员 2 的评分变化，重复上述过程并计算 $r_2^{'}=1629$。在这个例子中，球员 1 战胜了球员 2，尽管如此，球员 1 并没有获得那么多的评分，因为他的评分可靠性非常客观。 另一方面，球员 2 的评分发生了巨大变化，因为他的评分不可靠，相较于球员 1，他参加的比赛很少。

4.3 初始马尔可夫模型

在本节中，基于改进的 Glicko 排名系统估计发球的概率，构建初始马尔可夫模型，并对改进的 Glicko 排名系统进行参数优化。首先从如何选择模型比较标准开始。

4.3.1 模型评估指标

为了优化模型中的参数并比较不同的模型，需要一个评估指标。一个好的度量应该易于计算，并且在每个模型中都可用。这里使用 Glickman[25]和 Alex Mortan[26]

使用的度量：

$$d_{ij} = -s_{ij}\log\left(p_{ij}\right) - \left(1-s_{ij}\right)\log\left(1-p_{ij}\right) \tag{4.24}$$

其中d_{ij}表示球员j在第i场比赛中的误差、s_{ij}表示球员j在第i场比赛中的得分、p_{ij}表示球员j在第i场比赛中的估计得分。模型d的误差计算为所有d_{ij}的和。

使用上述模型误差来优化模型参数。此外还计算了其他两个模型强度预测指标。第一个是模型计算的正确预测的百分比。如果模型预测比赛实际获胜者赢得比赛的概率超过 50%，则该模型已计算出对比赛的正确预测。第二个评估标准是模型对投注交易所的表现如何。在这里，我们从四大交易所中获得最佳赔率。不幸的是早期的数据丢失了一些赔率数据。

为了对评估工作有一个基本的了解，评估了一个完全随机的模型，它返回一个从 0 到 1 的概率值来预测每场比赛。表 4.2 为随机模型的评估。它给出了模型的训练误差、偏差和均方误差。表 4.3 给出了 2016 年、2017 年、2018 年和 2019 年的正确预测百分比和投注的投资回报率。

表 4.2 随机模型评估

Model name	Error	Bias	Sqme
Random	27996.43889	0.49867	0.33102

表 4.3 随机模型正确预测百分比与 ROI

Year	Correct predict percentage	Betting ROI
2016	0.49164	-0.12814
2017	0.50727	0.05381
2018	0.49609	-0.02936
2019	0.49212	-0.03105

随机模型的大多数结果都是符合预期的。该模型的偏差和正确预测的百分比接近 0.50。更有趣的是投资回报率。欧洲博彩业的利润率一般在 10%左右，随机模型在 2016、2018、2019 年的平均投资回报率接近博彩业利润率，但 2017 年的回报率为 5%，这是非常成功的一年。但是在评估投资回报率的时，应该谨慎。尽管如此，仍将展示模型的投注结果，因为这是一个非常通用的指标，通常用于评估网球模型。

4.3.2 模型

使用前面几节中的思想，基于改进的 Glicko 排名系统构建初始马尔可夫模型。该模型使用改进的 Glicko 排名系统来估计球员在发球时赢得一分的概率。它将这些概率输入到马尔可夫模型中，以获得球员赢得比赛的概率。

在第 4.1 节中，详细介绍了如何计算一名球员赢得比赛的概率，给出了球员在发球中赢得一分的概率—表示为 p 和 q 。

在我们的模型中，将网球比赛视为两场单独的比赛：球员 1 发球对阵球员 2，球员 2 发球对阵球员 1。w_1表示球员 1 发球得分的百分比，w_2表示球员 2 发球得分的百分比。然后为每个球员分配两个等级，即发球评分等级和接发球评分等级。对于球员 1，将发球评分等级表示为r_1^s，将接发球评分等级表示为r_1^r。此外，每个球员都被分配了一个评分可靠性RD。

在第 4.2.3 节中，给出了改进后的 Glicko 排名系统的布局方程。使用其中两个，如等式 4.25 所示，现在可以计算概率 p 和 q 。

$$\begin{aligned} E\left(s|\ r_1, r_2, RD_2\right) &= \frac{1}{1+10^{-g\left(RD_2\right)\left(r_1-r_2\right)/400}} \\ g(RD) &= \frac{1}{\sqrt{1+3q^2\left(RD^2\right)/\pi^2}} \end{aligned} \tag{4.25}$$

要计算p，必须使用评分等级r_1^s和r_2^r，计算q使用评分等级r_2^s和r_1^r。由此推导出以下等式：

$$\begin{aligned} p &= E\left(s|\ r_1^s, r_2^r, RD_2\right) \\ q &= E\left(s|\ r_2^s, r_1^r, RD_1\right) \end{aligned} \tag{4.26}$$

然后可以将p和q值输入到马尔可夫链模型中，并计算球员赢得比赛的概率。比赛结束后，可以使用球员发球得分的实际概率来更新两名球员的评分。

以兹维列夫和费德勒为例。兹维列夫发球评分等级为$r_s^1 = 1504$，接发球评分等级为$r_r^1 = 1434$，费德勒发球评分等级为$r_s^2 = 1583$，接发球评分等级为$r_r^2 = 1491$。为两个球员设置了相同的评级可靠性$RD_1 = RD_2 = 100$。如果这两个球员互相比赛，可以使用等式 4.27 计算预期的p和q。

$$
\begin{aligned}
g(RD) &= \frac{1}{\sqrt{1+3q^2\left(RD^2\right)/\pi^2}} \\
&= \frac{1}{\sqrt{1+3*0.005^2\left(100^2\right)/\pi^2}} \\
&= 0.96404 \\
p &= E\left(s|\ r_1^s, r_2^r, RD_2\right) \\
&= \frac{1}{1+10^{-g(RD_2)\left(r_1^s-r_2^r\right)/400}} \\
&= \frac{1}{1+10^{0.96404*(1504-1491)/400}} \\
&= 0.518 \\
q &= E\left(s|\ r_2^s, r_1^r, RD_1\right) \\
&= 0.695
\end{aligned}
\tag{4.27}
$$

然后将计算得到的值$p = 0.518$和$q = 0.695$输入第 4.1.4 节中的马尔可夫链模型。兹维列夫赢得三局两胜制比赛的计算概率仅为$P^M = 0.011$。这个预测结果与实际的比赛结果相符。如果要对球员下注，应该把赌注押在费德勒身上。

比赛结束后，观察给出的统计数据。兹维列夫在发球的 100 分中赢得了 45 分。这意味着$p_{true} = 0.45$。使用这个值来更新球员的排名。详情请参阅第 4.3.2 节中的示例。

4.3.3 参数选择

由于模型的基础是一个排名系统，所以需要选择表现最好的一个进行进一步分析。使用第 4.3.1 小节中介绍的误差度量和 Nelder Mead 最小化方法优化参数。

要使用改进的 Glicko 排名系统，必须优化其参数：RD_{begin}、RD_{end}和RD_{step}。接下来首先估算参数RD，这些参数与 Glicko 例子中使用的参数类似，并且在 20 场比赛后稳定评级。因此，选择的初始参数是$RD_{begin} = 300$、$RD_{end} = 100$ 和$RD_{step} = 10$。这将是进行优化的起点。

使用第 4.3.1 小节中介绍的误差度量和 Nelder Mead 最小化方法对参数进行优化。在表 4.4 中给出了初始值及其收敛值。

表 4.4 初始马尔可夫模型参数优化的初始值和收敛值

初始值			收敛值			误差
RD_{begin}	RD_{end}	RD_{step}	RD_{begin}	RD_{end}	RD_{step}	
300.0	100.0	10.0	116.5	45.8	19.0	21171.1
90.0	45.0	15.0	55.0	45.7	19.6	21171.0
55.0	45.0	10.0	55.0	45.7	10.1	21171.0
30.0	10.0	4.0	32.5	45.8	10.17	21171.7

参数从大多数点收敛到$RD_{begin} = 55.0, RD_{end} = 45.7$，$RD_{step} \geq 10$。因为$RD$在 20 场比赛后变得比$RD_{end}$小，所以$RD_{step}$的值是多少并不重要。模型还可收敛到局部最优 $RD_{begin} \leq 45.8, RD_{end} = 45.8$。对于这些值，排名系统是 Elo 排名系统。观察发现，添加 RD 变量只有很小的优势。在接下来改进的模型中，将使用 Elo 排名系统代替改进的 Glicko 排名系统。

表 4.5 初始马尔可夫模型评估

Model name	Error	Bias	Sqme
Base Markov	21171.03847	0.40178	0.21011

表 4.6 初始马尔可夫模型正确预测百分比与 ROI

Year	Correct predict percentage	Betting ROI
2016	0.6945	0.04011
2017	0.67343	-0.01659
2018	0.68019	0.04691
2019	0.67866	-0.01552

表 4.5 给出了初始马尔可夫模型在$RD_{step} = 10$，$RD_{begin} = 55.0$，$RD_{end} = 45.7$时的训练误差、偏差和均方误差。表 4.6 给出了初始马尔可夫模型在$RD_{step} = 10$，$RD_{begin} = 55.0$，$RD_{end} = 45.7$ 时，对于 2016-2019 年的网球赛事的正确预测百分比和投注的投资回报率。观察发现，初始马尔可夫模型正确预测百分比的均值为 0.68170，偏差接近 0.4。对网球比赛投注的平均投资回报率为 1.7%，其中 2016 年和 2018 年的投资回报率都达到了 4%。

4.3.4 图形工具

上面所提出的模型可以用于衡量球员状态。每个球员都从两个方面被评估，发球表现和接发球表现。这些可以表示为二维平面中的点。为此创建了一个 web 应用程序，以图形化的方式展示球员的发球与接发球等级排名是如何随时间变化的。该图形工具的外观如图 4.3 所示。

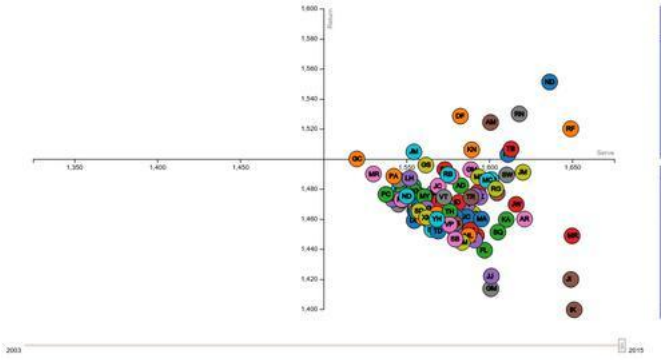

图 4.3 球员的发球与接发球等级

该图形工具使用 2003 年到 2019 年的数据。为了更清晰的呈现，值是在逐个锦标赛的基础上选择的，并使用指数函数进行平滑处理。所创建的图形工具用于发现现有模型中的问题。

4.4 初始马尔可夫模型的问题及解决方案

初始马尔可夫模型并未将球员发球与接发球的内在差异、场地类型的影响以及球员发展势头这三个变量纳入模型。本节对未纳入变量的影响进行分析，并提出相应的解决方案。

4.4.1 发球与接发球差异

首先观察分析在初始马尔可夫模型中，球员在自己的发球局中赢得一分（p,q）的预期概率与球员在发球局中赢得一分的实际概率。在图 4.4 中，对这些概率进行了比较。

模型对于球员发球得分的概率的平均预测值为 0.6087，而球员实际发球得分

的平均值为 0.6312，二者的平均差值是 0.0224，这意味着初始马尔可夫模型低估了球员在自己的发球局上赢得一分的概率。预测值与实际值之差的方差为 0.0895。

排名数据也存在偏差。图 4.5 显示了图形工具的截图，显示了 2018 年温网排名前 100 名的球员的发球与接发球等级。从图 4.5 可以看出，球员的发球评分等级通常大于接发球评分等级。

图 4.5 所示的结果是符合预期的，因为发球时球员更具有优势。所以，这就是为什么初始马尔可夫模型总是低估球员在发球时赢得一分的概率。

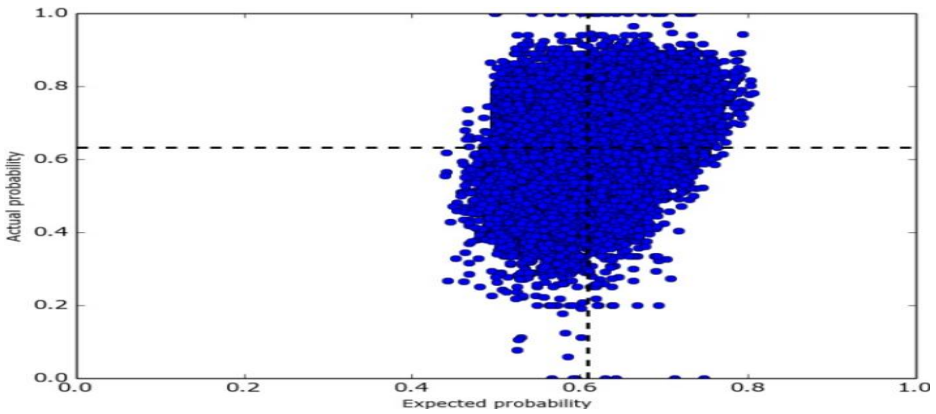

图 4.4 球员在发球局中赢得一分的预期与实际概率

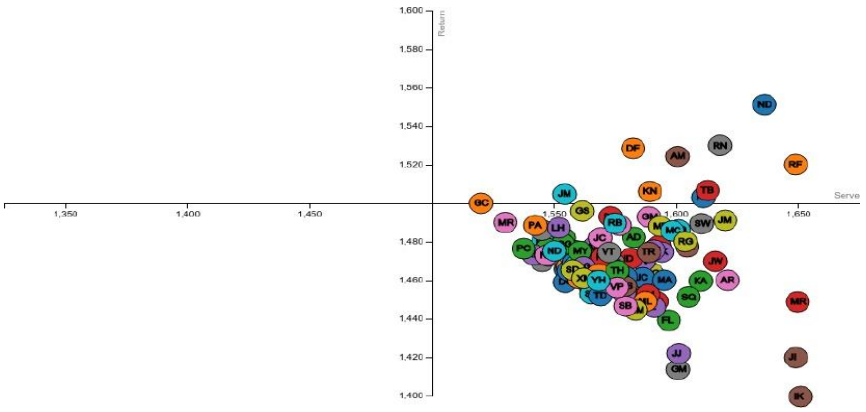

图 4.5 2018 温网排名前 100 名的球员的发球与接发球等级

上一节提出的初始马尔可夫模型出现了一些问题，因为发球得分概率的预测值和实际值的平均值并不相同。

$$\begin{aligned} p &= E\left(s|\ r_1^s, r_2^r, RD_2\right) \\ q &= E\left(s|\ r_2^s, r_1^r, RD_1\right) \end{aligned} \tag{4.28}$$

由于上述公式无法返回正确的结果，因此应相应修改。解决方案是增加另一个变量，这意味着发球与接发球之间存在内在差异。为此将方程改写为：

$$\begin{aligned} p &= E\left(s|\ r_1^s, r_2^r, RD_2\right) + e \\ q &= E\left(s|\ r_2^s, r_1^r, RD_1\right) + e \end{aligned} \tag{4.29}$$

在计算p和q值时应该加上e。然而，在更新评级时，还必须从实际结果中减去e。

对于$e = 0.1$，创建一个散点图并计算平均值。计算的平均值现在为 0.626，与实际平均值相差 0.005。此外，方差已降至 0.0835。新结果如图 4.6 所示。

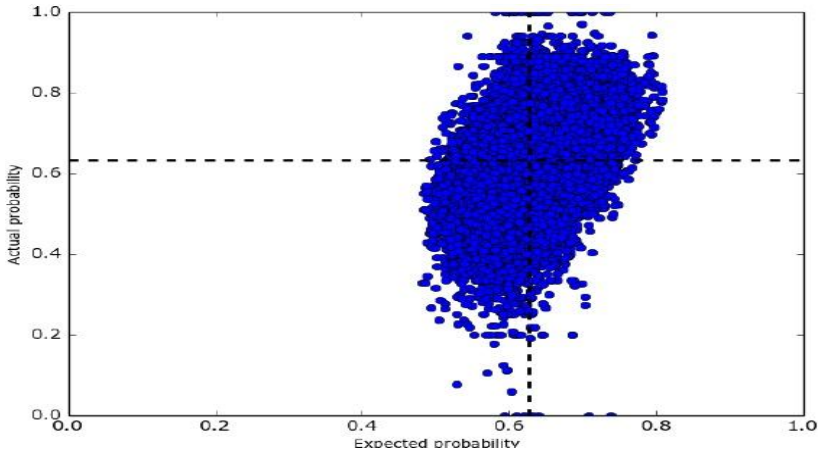

图 4.6 修正后球员在发球局中赢得一分的预期概率与实际概率

4.4.2 场地类型的影响

到目前为止，本文忽略了比赛分析的一个重要方面，那就是场地类型。每种不同的场地类型都有影响比赛风格的特点，网球赛事尤其不同，它有硬地、草地、红土及地毯四种完全不同类型的场地。有些球员在不同的赛场上表现完全不同。

上述的初始马尔可夫模型完全忽略了场地类型的影响，这是一个很大的错误。在我们的建模方法中，没有直接的方法将其纳入模型中。在本小节中，将着重探讨这种影响，以及如何使用它来改进前面提出的初始马尔可夫模型。

（1） 场地类型比较

不同类型的场地之间最重要的区别是它们的硬度不同，硬度会影响球员发球的力量及强度。场地球速越快，球员在自己发球局赢得一分的概率也就越高。图 4.7 比较了球员在不同场地上发球得分的概率。红土场，作为球速最慢的场地，球员

在该种场地发球得分的平均概率为 0.611，在硬地的平均概率为 0.637，地毯场的平均概率为 0.653 以及草地场的平均概率为 0.657。该图展示了球员发球得分概率从球速最慢的场地到球速最快的场地是越来越高的。

直觉上的猜测是，具有更高发球评分等级的球员在像草地一样的快速场地上具有更多的优势。下面用图 4.8 和 4.9 来证实上述假设。图 4.8 和 4.9 表明：发球评分等级从球速最慢的场地到球速最快的场地呈现出由慢到快的上升趋势，而接发球评分等级则呈现出下降的趋势。

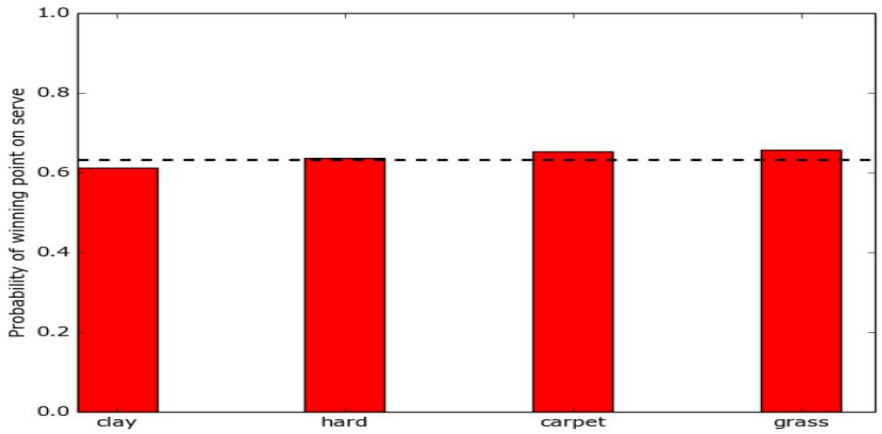

图 4.7 球员在不同场地上发球得分的概率

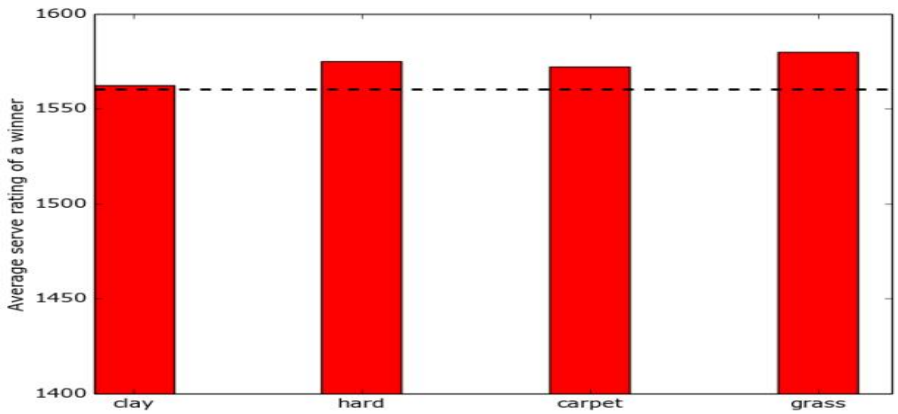

图 4.8 不同场地上获胜者的平均发球评分等级

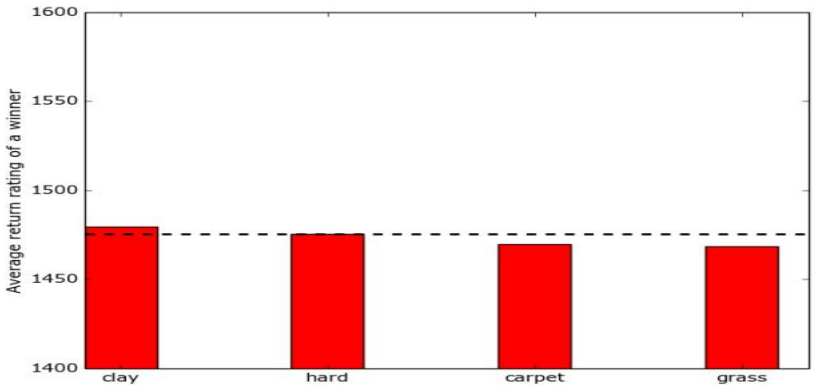

图 4.9 获胜者在不同场地上的平均接发球评分等级

（2） 案例研究

选择安迪·穆雷作为研究对象，研究不同的场地是如何影响其表现的。在数据库中，他的比赛场次为 662 场。在表 4.7 中可以得到安迪·穆雷在不同类型的场地比赛的统计数据。

表 4.7 安迪穆雷在不同场地类型的统计数据

场地类型	比赛胜率	发球得分概率	接发球得分概率	比赛数量
硬地	0.7748	0.6558	0.4305	444
红土	0.6636	0.6236	0.4320	107
草地	0.8300	0.7010	0.4003	100
地毯	0.7273	0.6509	0.4001	11
全部	0.7644	0.6573	0.4257	662

从表 4.7 中的数据可以看出，安迪穆雷在草地上表现最为出色，而他表现最薄弱的场地是红土场。安迪穆雷在草地上比赛时，发球得分概率可以达到 0.7010，这比他的平均水平高出 0.0437。这是应该能够被纳入初始马尔可夫模型中的东西。

继续观察穆雷在草地上的表现。根据正态分布对他发球局得分的概率进行建模，如图 4.10 所示。虽然正态分布拟合数据较好，但一个标准差区间非常的宽，甚至包含了总体均值。根据正态分布，穆雷在草地上的平均发球表现实际上优于他的平均发球表现的概率不到 68%。这是相当令人惊讶的，因为研究使用了大量的安迪·穆雷的比赛数据进行分析。

接着用伯努利分布对他发球局得分的概率进行建模。如图 4.11，在这样的数据集上，一个标准差区间现在比预期的要小得多。在初始马尔可夫模型中增加场

地类型的影响的一个非常简单的想法是简单地调整球员在不同场地发球或接发球评分等级，这在统计上是显著的。所以，不应忽略不同场地之间的内在差异。

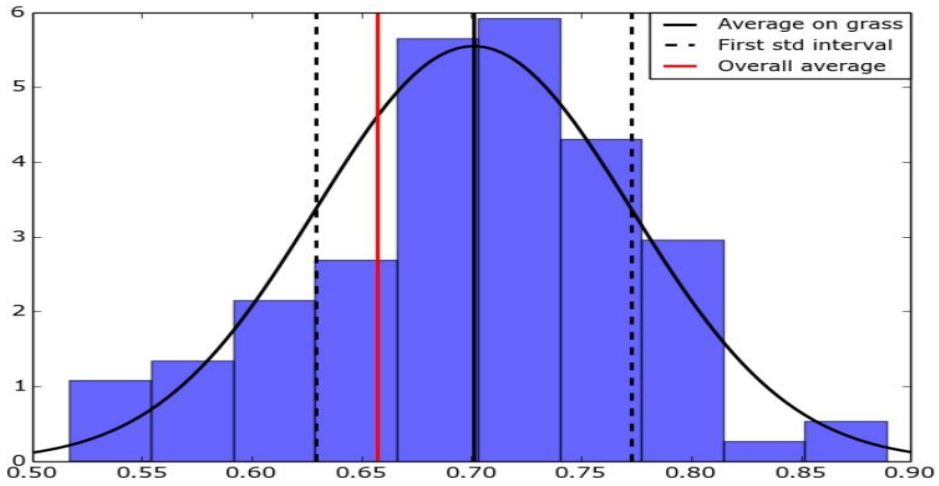

图 4.10 安迪穆雷在草地上发球得分的概率分布，使用正态分布对数据进行建模。

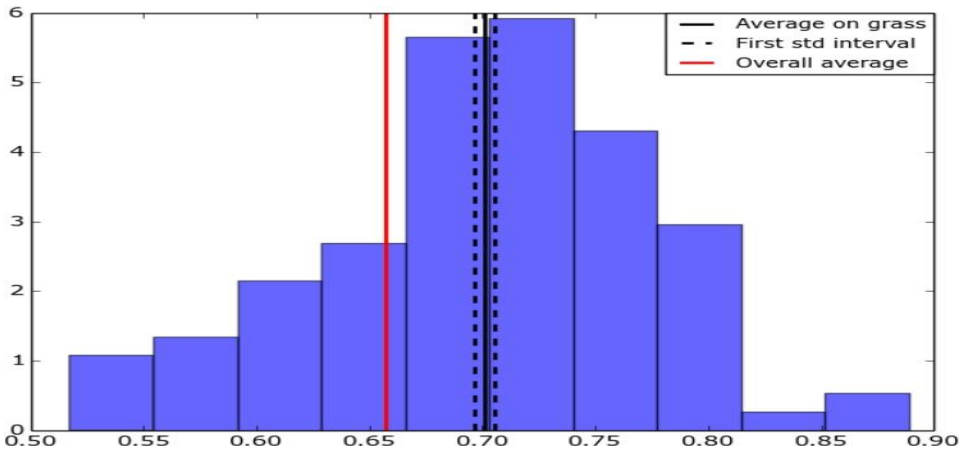

图 4.11 安迪穆雷在草地上发球得分的概率分布，使用伯努利分布对数据建模

（3） 解决方案

在前面观察的基础上，尝试将场地类型影响添加到公式中。对球员在自己发球局赢得一分的概率 (p,q) 进行修正，公式如下：

$$p' = p + p_\alpha \tag{4.30}$$

为了计算 p_α，使用本小节后面介绍的一个算法。首先，需要计算输入变量 $input$ 和 h：

$$
\begin{aligned}
s_1 &= s_1^{\text{surface}} - s_1^{\text{all}} \\
r_2 &= r_2^{\text{surface}} - r_2^{\text{all}} \\
input &= s_1 - r_2 \\
h_1 &= \sqrt{\frac{s_1^{\text{surface}}\left(1 - s_1^{\text{surface}}\right)}{100 * n_1^{\text{surface}}}} * z_{1-\alpha/2} \\
h_2 &= \sqrt{\frac{r_2^{\text{surface}}\left(1 - r_2^{\text{surface}}\right)}{100 * n_2^{\text{surface}}}} * z_{1-\alpha/2} \\
h &= h_1 + h_2
\end{aligned}
\tag{4.31}
$$

s_1^{surface} -球员 1 在特定场地类型发球局得分的平均概率

s_1^{all} -球员 1 在所有场地发球局中的平均得分概率

r_2^{surface} -球员 2 在特定场地类型接发球局得分的平均概率

r_2^{all} -球员 2 在所有场地发球局中的平均得分概率

n_1^{surface} -球员 1 在特定场地上进行的比赛场次

n_2^{surface} -球员 2 在特定场地上进行的比赛场次

α-置信系数

使用如下算法得到p_α值：

算法 1：计算场地类型影响

1：if $input + h < 0$:$p_\alpha = input + h$

2：elif $input + h > 0$:$p_\alpha = input - h$

3：else $p_\alpha = 0$

4：return p_α

5：end

4.4.3 球员发展势头

初始马尔可夫模型缺乏的最后一个相关的变量，是球员的上升势头。在这个模型中，没有变量可以跟踪球员表现的波动性。如果一名球员的水平突然开始快速提升，系统就需要花一些时间才能更新最新的情况。这个问题也可以看作是球员的状态预测。在这一小节中，着重分析我们的系统如何低估了日本球员锦织圭

的表现，并提出了模型在跟踪球员势头方面可能的改进。

（1）案例研究

通过一个例子来说明初始马尔可夫模型缺乏球员发展势头变量。从 2011 年到 2015 年，日本球员锦织圭的水平一直在稳步攀升，尤其是 2014 年，他的表现异军突起，取得了很大的突破[37]。在他参加 2015 年温布尔顿网球公开赛时，他在 ATP 官方排名中排名第四，而我们的模型将他排在第十位（如果根据发球和接发球评分等级之和对球员进行排名）。

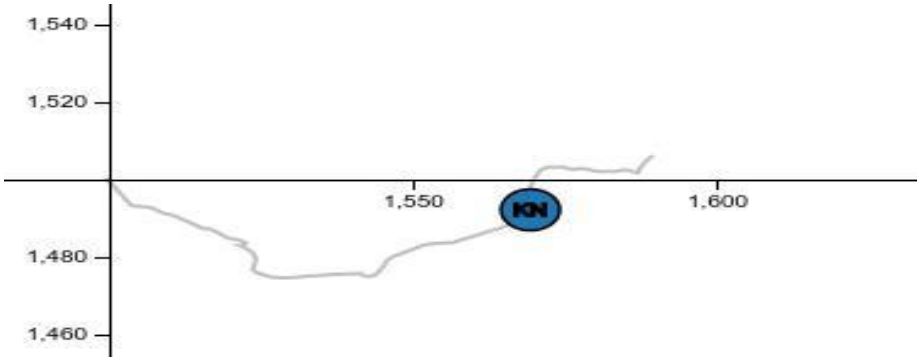

图 4.12 锦织圭的发球与接发球等级

图 4.12 显示的是一个图形工具中的屏幕截图，该屏幕截图表现了锦织圭从 2011-2015 年的进步。圆圈设置在 2012 年年中，可以看到他的稳步攀升和 2014-2015 年水平的飞跃。图 4.13 比较了锦织圭的 ATP 排名和他在初始马尔可夫模型中的排名。初始马尔可夫模型是跟踪 ATP 排名，但有一个延迟，因为 Elo 排名系统需要一段时间才能收敛到实际排名。

（2）解决方案

对于手头的问题，有两个很好的解决方案。 但是，它们都不能成功地包含到初始马尔可夫模型中。

首先，Glicko 的排名系统 Glicko2 应该是解决该问题所需要的。它包括一个跟踪球员排名波动的变量。如果球员开始提高的更快，他的评级将提高更大的幅度，直到稳定。但是没有成功地调整 Glicko2 排名系统以适应初始马尔可夫模型。

其次，Bester 进行了一些研究，为 Elo 排名系统增添球员势头变量。他提出了三种不同的方法来为 Elo 排名系统增加球员势头变量，但它们都没有改善模型的性能[38]。

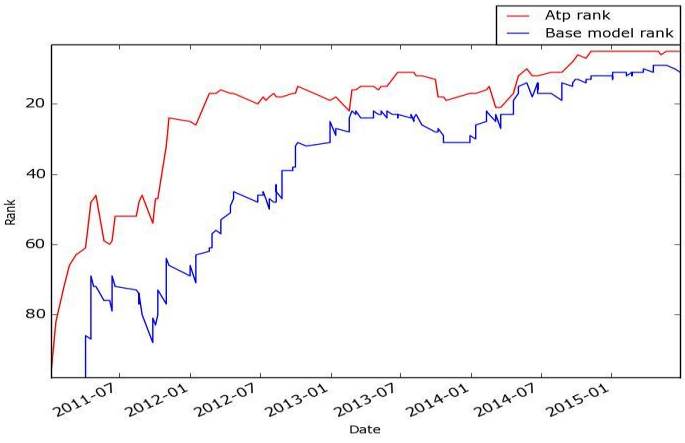

图 4.13 锦织圭的 atp 排名和基础模型排名的比较

4.5 改进的马尔可夫模型

本节使用前面几节的结果，对初始马尔可夫模型进行改进。改进的马尔可夫模型使用 Elo 排名系统，而不是改进的 Glicko 排名系统。此外，模型还添加了修正发球与接发球之间内在差异的边缘常数 e 与不同类型场地影响的参数，提高了初始马尔可夫模型预测的准确率。

4.5.1 模型

球员 1 的发球评分等级和接发球评分等级分别用 r_1^s 和 r_1^r 来表示，同样地，r_2^s 和 r_2^r 表示球员 2 的发球评分等级和接发球评分等级。首先回顾第 4.2.1 小节中预期得分的 Elo 排名公式：

$$E(r_1,r_2)=\frac{1}{1+10^{(r_1-r_2)/400}} \tag{4.32}$$

接下来使用以下两个等式计算球员在发球时赢得一分的概率：

$$\begin{aligned} p &= E\left(r_1^s,r_2^r\right)+0.1+p_\alpha \\ q &= E\left(r_2^s,r_1^r\right)+0.1+q_\alpha \end{aligned} \tag{4.33}$$

式中，p_α 和 q_α 为场地影响调整参数，将他们的初始值设置为 $p_\alpha=0$ 和 $q_\alpha=0$。每场比赛后，都会更新两名球员的评分等级。用 s_1 来表示球员 1 发球得分的百

分比，用s_2来表示球员 2 的发球得分百分比。

使用下列等式更新评分等级：

$$\begin{aligned} r_1^s &= K*\left(s_1 - E\left(r_1^s, r_2^r\right)\right) \\ r_1^r &= K*\left(E\left(r_2^s, r_1^r\right) - s_2\right) \\ r_2^s &= K*\left(s_2 - E\left(r_2^s, r_1^r\right)\right) \\ r_2^r &= K*\left(E\left(r_1^s, r_2^r\right) - s_1\right) \end{aligned} \tag{4.34}$$

由于目前将场地影响p_α和q_α设置为 0，所以优化的唯一参数是参数K。在下一小节中将展示如何对它进行优化。

4.5.2 参数优化

由于只有一个参数需要优化，下面将模型误差绘制成参数 K 的函数，如图 4.14 所示。

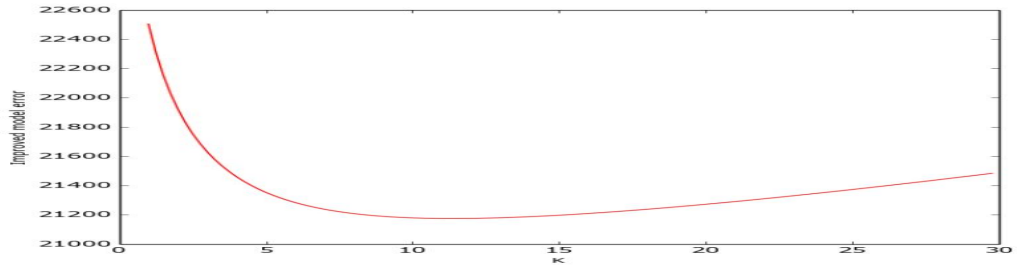

图 4.14 改进模型对不同 K 值的误差

参数 K 的函数为凸函数，在$K = 12$ 附近误差最小。对于大多数优化方法和起点，应该包括最优值。运行优化两次，从$K = 6$ 开始一次，从$K = 20$ 开始一次。在这两种情况下，参数都收敛到$K = 11.35$，误差为 21174.8。 模型的误差略大于使用修改后的 Glicko 排名系统的模型中的误差。关于 K 的三种取值的改进模型，均具有超过 66%的预测百分比。

4.5.3 增加场地影响

在第 4.4.2 节中，分析了不同类型的场地的影响并提出了可用于改进现有模型的方程。场地影响只有一个参数，即α。在改进的模型中添加场地影响时，使用以

下两个方程计算p和q：

$$\begin{aligned} p &= E\left(r_1^s, r_2^r\right) + 0.1 + p_\alpha \\ q &= E\left(r_2^s, r_1^r\right) + 0.1 + q_\alpha \end{aligned} \tag{4.35}$$

式中，p_α和q_α按照算法 1 计算。

测试不同α取值下改进的马尔可夫模型的误差，结果见表 4.8。正如预期的那样，对于$\alpha = 1.0$ 的模型，没有考虑场地类型的影响。对于大于 0.5 的α，改进的模型表现比没有考虑场地类型影响时要好。对于小于 0.5 的α，改进的模型表现比没有考虑场地类型影响时更差。

表 4. 8 具有场地类型影响的改进的马尔可夫模型，不同α取值下的误差

α	改进模型误差
0.25	21401.2
0.5	21261.3
0.68	21172.6
0.85	21091.7
0.95	21051.0
0.975	21043.1
0.9875	21043.9
0.9925	21074.0
1.0	21174.8

表 4.9 给出了改进的马尔可夫模型在$K = 11.35$，$\alpha = 0.975$ 时的训练误差、偏差和均方误差。表 4.10 给出了改进的马尔可夫模型在$K = 11.35$，$\alpha = 0.975$ 时，对于 2016 年、2017 年、2018 年和 2019 年的网球赛事的正确预测百分比和投注的投资回报率。观察发现，改进的马尔可夫模型正确预测百分比的均值为 0.68822，偏差为 0.399。对网球比赛投注的平均投资回报率为 0.9%，其中 2016 年的投资回报率都达到了 5%。比起初始马尔可夫模型，正确预测的百分比提高了 0.65%，偏差也有所下降。

表 4.9 改进的马尔可夫模型评估

Model name	Error	Bias	Sqme
Improved Markov	21043.60624	0.39916	0.20856

表 4.10 改进的马尔可夫模型正确预测百分比与 ROI

Year	Correct predict percentage	Betting ROI
2016	0.69704	0.05207
2017	0.68377	-0.01839
2018	0.68059	-0.00437
2019	0.69146	0.00905

4.6 改进的马尔可夫模型的应用

本节主要介绍改进的马尔可夫模型可以应用于指导球员训练与预测比赛时长两个主要领域。对于每个领域，都给出一个示例来演示模型是如何应用的。

4.6.1 指导球员训练

改进后的模型与 4.3.4 节提供的图形工具相结合，可用于指导球员的训练。如果球员的接发球排名有所提高，但发球排名没有提高，教练可以想办法调整球员的训练，包括增加更多的发球训练和发球比赛。发球排名高于接发球排名的球员则相反。

在图 4.15 中，可以看到伊沃卡洛维奇的发球与接发球表现。他有非常强势的发球表现，水平接近世界最顶尖的球员诺瓦克德约科维奇，但他的接发球表现相当差。虽然他肯定有受过专门的训练来纠正这一点，但通过模型中使用的类似排名方法可以为他和他的教练团队提供更详细的信息。

4.5.2 小节对模型的参数进行了优化，以使模型的度量误差最小化，参数的优化可以根据使用模型应用的具体方面进行调整。例如在训练一名球员时，可以将参数K设置得很高，并根据训练内容的改变在短时间内观察球员发球与接发球评级的变化。虽然该模型的当前状态不会为训练球员提供非常明确实际的建议，然而，我们相信基于不同数据的排名方法能够帮助教练发现所执教球员的弱点。在球员及教练并没有过多关注的一些技术缺陷方面，使用模型可能非常有用。

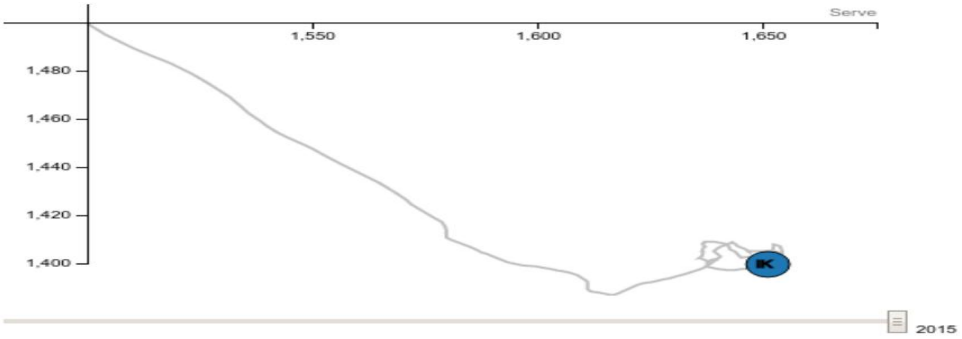

图 4.15 伊沃卡洛维奇的发球与接发球表现

4.6.2 比赛时长预测

如前所述，媒体对估计网球比赛的持续时间非常感兴趣。由于每场比赛的比赛时长都不是固定的，因此必须估计媒体日程。根据这方面的需求，提出了一种使用蒙特卡罗模拟和正态分布拟合来估计网球比赛持续时间的简单方法。

首先比较以分钟为单位的比赛时长与一场比赛的总得分数。图 4.16 表明二者存在线性关系。如果将时长（以分钟为单位）表示为m，将得分表示为p，则可以写出如下的线性方程：

$$m = 0.6756 * p - 1.257 \tag{4.36}$$

通过使用蒙特卡洛模拟，可以找到得分的分布，使用上述等式将其转换为比赛持续时间，然后拟合正态分布。因此得到的正态分布可用于得到比赛的预期长度，但更重要的是还可以获得置信区间。

接下来选择 2019 年温网决赛来测试上面的方法，在 2 小时 56 分钟的激烈的网球比赛中，诺瓦克德约科维奇击败了罗杰费德勒。改进的马尔可夫模型估计了在二人的比赛中，诺瓦克德约科维奇发球局赢得一分的概率为$p = 0.666$，罗杰费德勒在自己发球局赢得一分的概率为$q = 0.640$。

在 Python 中实现的蒙特卡洛模型中输入p和q值，以获得得分的分布。使用方程 4.36 将这些值转换为比赛时长，并最终将正态分布拟合到数据中，这样就可以得到图 4.17 所示的分布。使用模型估计这场决赛的持续时间为 179 分钟，这是一个精确的估计，与实际比赛时长只差 3 分钟。

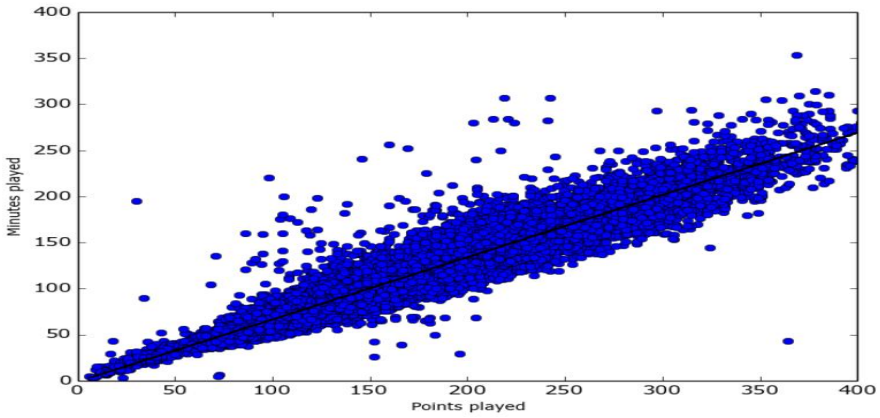

图 4.16 比赛时长与得分相关性

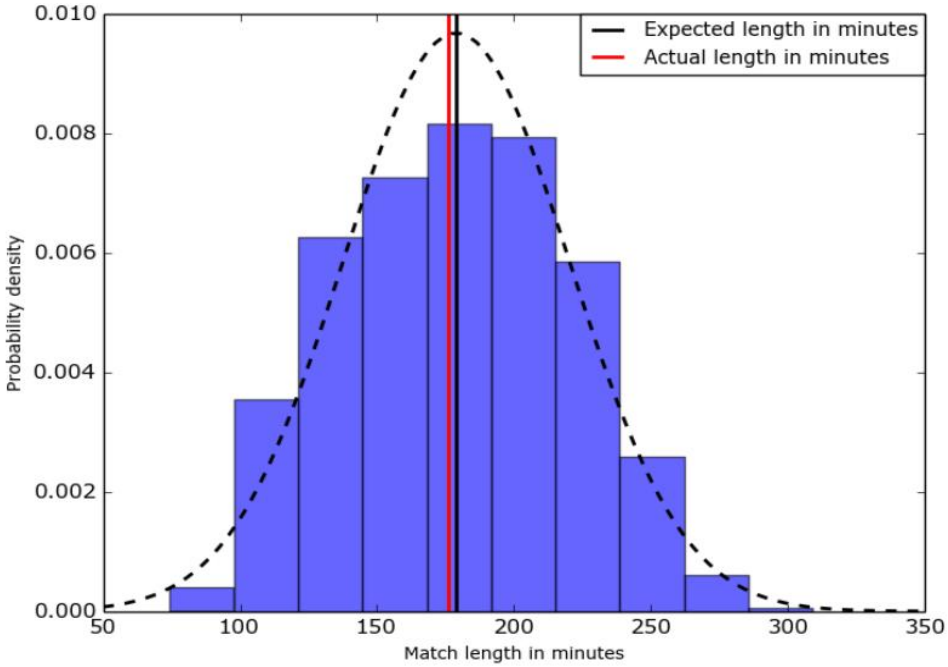

图 4.17 2019 年温网决赛的比赛时长分布（以分钟为单位）

第五章 基于机器学习模型的网球赛果预测及投注

本章基于 2000 年到 2019 年的 ATP 比赛的所有数据，使用机器学习中的支持向量机模型、逻辑回归模型、以及多层感知机模型对网球赛果进行预测，将机器学习模型结果与使用经验法则的预测结果进行比较，并进行特征分析，最后使用 xgboost 模型和赔率数据进行网球投注。

5.1 数据来源与处理

该部分研究使用由 Jeff Sackmann 创建的存储库中的网球比赛数据，该存储库包含了从 1968 年到 2021 年的所有 ATP 比赛[39]。在每场比赛中，详细统计了胜负双方的历史交手记录、胜率、世界排名、ace 数、双误数、一发成功率、一发得分率、二发得分率、面临破发点数以及挽救破发点个数等技术统计数据，这些统计数据提供了大量的关于网球比赛的信息。以 2000-2019 年的 44856 场比赛数据为样本数据，机器学习分类器可以根据这些信息进行训练。

首先，检查原始数据中每个特征的特殊值，并删除在特征提取阶段无法使用该值的数据。最主要的就是一些比赛因为球员退赛导致的缺失值，虽然考虑了如何进行缺失值的处理，但是插补缺失数据可能会改变分类的结果。因此，最终决定删除所有具有缺失值的数据。然后从数据集中删除所有的戴维斯杯比赛及所有的非 ATP 官方的比赛。

最后，该研究将原始数据集转化为一个二分类问题。原始数据集包含两名球员的技术统计数据和获胜者的名字，这不是一种容易处理的数据格式。所以，将选手标记为 “球员 1”和“球员 2”，然后添加一个标签列，如果“球员 1”赢得比赛，数值将为 1，否则为 0。

5.2 特征提取

该部分使用了三种不同的方法来提取数据的特征。在提取所有特征之后，创建一个长度等于特征数量的特征向量，并将前面提到的球员 1 的输赢作为预测的标签。下面对特征的提取方法进行了总结。

5.2.1 特征对称

在原始数据中，每条数据均为每个统计量提供两个值。例如，选手 1 和选手 2 的“aces”统计数据都被记录下来。这种表示方式虽然提供了每个选手的技术统计信息，但它也有一个缺点；如果研究中交换选手 1 和选手 2 的标签，那么构建的分类器就会给每个特征赋予不同的权重，从而预测不同的结果。为了避免这种情况，可以使用原始统计数据创建一个单一特征，如以下公式所示。

$$Feature_i = Raw_{1,i} - Raw_{2,i} \tag{5.1}$$

其中 i 是示例索引，Raw 是来自原始数据集的原始特征，Feature 是新生成的特征。Clark 和 Dyte 之前的工作表明这种表示是有效的，所以以这种方式转换所有的原始特征[40]。

5.2.2 共同对手模型

这一部分使用 Knottenbelt 的共同对手模型[34]。在这个模型中，首先考虑球员 1 和球员 2 都遇到过的一组共同对手。对于每个球员 1 和 2，可以找到针对一组共同对手中每个球员的统计数据的平均值。最后，使用这两个平均值创建一个差异变量。

这一模型背后的理念是，当根据共同的对手进行判断时，比较两名选手之间的统计数据将变得更有意义。因为模型考虑了每个选手面对不同对手的表现，所以能够更好地了解每个选手数据的真实价值。模型中所有的数值特征都是以这种方式计算出来的。

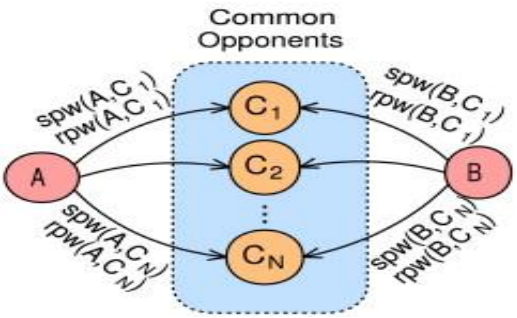

图 5.1 共同对手模型

5.2.3 派生特征

这一部分创造了派生特征作为所拥有的其他特征的函数。该派生特征也可被称为球员“全面性”特征。一直以来，在关于谁才是历史上最好网球球员的争论中，网球评论家们要求球员要有好的发球、精准的落点、好的接发球和良好的跑动，以及其他一些技术特点。拥有所有这些技术的球员就被认为是一个“全面的”球员。接下来试图构造评估球员全面性的特征，取一发得分率、二发得分率和成功挽救破发点百分比的乘积。然而，现有的数据集没有关于球员击球落点和球场跑动的信息，所以将该特征限制在三个组成部分。

$$Completeness_i = 1stWon_i * 2ndWon_i * \frac{bpSaved_i}{bpFaced_i} \tag{5.2}$$

通过将各部分技术指标相乘，便形成了一个派生特征，该特征要求选手在这些方面都能够有良好的表现。但是，如果缺少任何特征值，那么特征值就会下降，但可能不会对最终的预测产生太大的影响。初步认为，这些基于特定领域的特性提供了有关问题的额外信息，并可以增强模型的准确性。

5.3 模型

在进行分析时，使用通过共同对手模型预处理的训练数据进行训练。为了优化机器学习模型，我们评估了验证数据的性能。当然，不能使用与训练数据相同的方法进行预处理的验证数据进行预测，因为这将为分类器提供尚未发生的比赛信息。

5.3.1 支持向量机参数选择

使用以下方程中的软间隔支持向量机公式来构建分类器，其中ω为权重向量，y_i是示例 i 的标签，x_i是第 i 个示例，而C是一个超参数，也被称为惩罚参数，C 的取值越大，对错分类样本的惩罚就越大。

$$\min_{\omega} \frac{1}{2}\omega^T\omega + C\sum_i \max\left(0, 1 - y_i\omega^T x_i\right) \tag{5.3}$$

为了找到支持向量机的最优超参数，分别在三个核函数(线性、多项式和径向基函数)和惩罚参数 C 上使用网格搜索。在训练集上训练，在验证集上测试，给出

了表 5.1 的结果。从这个网格搜索实验中，可以发现 C = 0.5 的径向基核函数是支持向量机模型的最优超参数对。

表 5.1 调整 SVM 超参数的结果

Kernel	C	Accuracy
Linear	0.25	0.686
Linear	0.5	0.708
Linear	1.0	0.726
Linear	10	0.717
Polynomial	0.25	0.691
Polynomial	0.5	0.623
Polynomial	1.0	0.663
Polynomial	10	0.719
RBF	0.25	0.710
RBF	0.5	0.734
RBF	1.0	0.729
RBF	10	0.712

5.3.2 模型比较

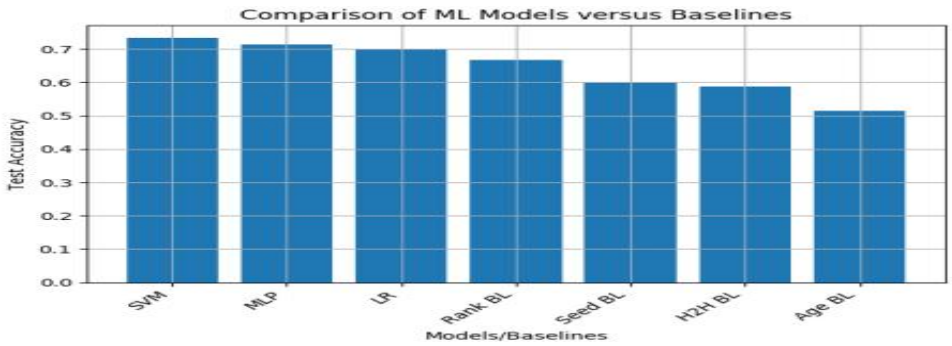

图 5.2 机器学习模型与经验法则预测对比

图 5.2 从左到右依次展示了使用支持向量机模型、多层感知机模型、逻辑回归模型、球员 ATP 排名、球员赛事排名、球员交手记录以及球员年龄对网球赛果预测的准确率。

从图 5.2，可以看出三种机器学习模型预测的准确率都超出了使用经验法则进行预测的准确率。基于经验法则进行预测的表现符合预期，无论是通过 ATP 排名、

赛事排名、交手记录还是球员年龄进行预测，准确率都超过 50%。三种机器学习模型的性能大致相同，预测准确率相差约 5%。其中支持向量机分类器的表现最好，预测准确率达到了 73%，但仍有改进空间。虽然多层感知器模型能够使用三层网络结构在一定的容错范围内逼近任何连续函数，但它的预测表现不如支持向量机模型，因为支持向量机不会产生局部最小值，会得到更少的错误分类。未来可以使用深度学习方法进一步开发多层感知机分类器，看看准确率是否会提高。综上，机器学习方法用于网球赛果预测可以获得更好的性能。

5.3.3 特征分析

为了找出决定比赛结果的最重要的特征是什么，特别进行了消融研究。在这样的研究中，首先得到包含了所有特征的分类器的准确率。然后，每次删除一个特征，以检查每个特征对模型的最终准确度的贡献。对所有的模型都进行了特征消融，其中性能最好的支持向量机分类器的消融研究结果如图 5.3 所示。

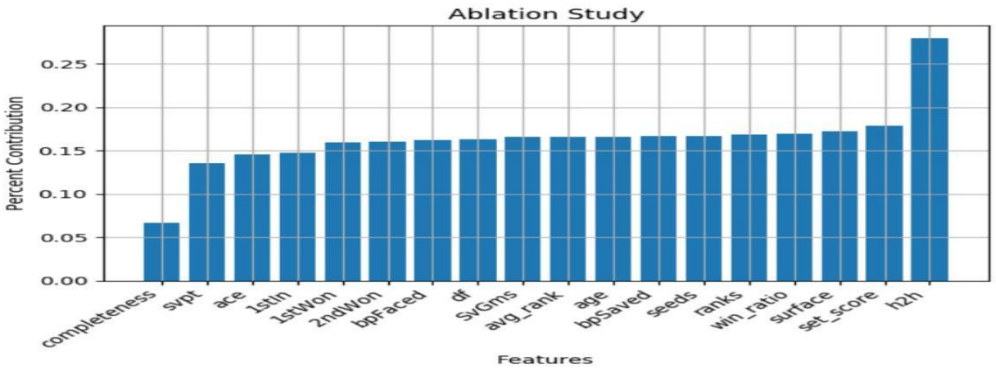

图 5.3 支持向量机的特征消融研究

通过观察，球员历史交手记录特征对测试准确性的贡献最大，达到了 28%。这意味着，一场比赛交手双方球员过去的比赛历史胜负记录比当前比赛中收集的任何统计数据都重要。接下来的两个最重要的特征是球员胜率和 Ace 球的数量。从直观上看，胜率(指的是球员在与所有其他球员，而不仅仅是当前对手对抗时获得的胜利)是衡量球员获胜几率的重要指标。但其他特征的贡献没有交手记录特征那么多，而且这些特征的贡献几乎是相同的。由此得出的结论是，该模型并没有提供到底哪些比赛技术统计，比如一发成功率或破发成功率，球员应该优先考虑

以获得比赛胜利。剩下的特征中值得注意的是球员全面性特征，这是前面创造的派生特征，用于衡量球员在比赛的多个方面的能力。该研究否定了网球评论员关于球员全面性的许多讨论。消融研究表明，每场比赛的统计数据对测试准确度都有一定的积极影响，决定比赛结果的最重要的特征是历史交手记录。为了对球员、教练、投注者提供有用的信息，表 5.3 总结了排名前五的特征及其贡献率。

表 5. 3 特征贡献

Feature	Percent Contribution
Head-to-Head	0.28
Win Ratio	0.169
Aces	0.168
Rank	0.168
Seed	0.166

5.4 网球投注

网球比赛赛果预测的一个非常流行的应用领域是网球投注。庄家和玩家都对预测网球比赛的结果很感兴趣。本节使用 xgboost 模型以及 Pinnacle 和 Bet365 两个博彩公司的赔率数据对网球比赛进行投注，估计不同投注策略的投资回报率（ROI）。首先从评估简单策略的投资回报率开始。

5.4.1 评估简单投注策略

（1）对所有比赛投注

首先对 2011 年到 2018 年之间的所有网球比赛下注。此时间段内的比赛场次为 16412 场。有这么多的比赛数据，置信区间会非常小。

接下来使用以下三种简单的投注策略：(a)总是对获胜赔率最小的球员进行投注；(b)总是在 ATP 排名更高的球员身上下注；(c)对每一场比赛完全随机投注。 假设在每场比赛中投注的钱都是一样的。在下面的图表中，通过两家博彩公司可以比较以上三种简单投注策略。

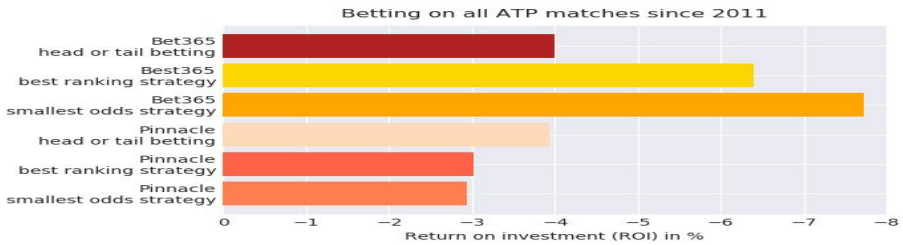

图 5.4 使用三种投注策略对所有比赛下注的投资回报率

由图 5.4 可以看出，如果对所有的比赛进行下注，那么不管使用以上哪种投注策略得到的投资收益率均为负。例如，如果在 1000 场比赛中，玩家通过 Pinnacle 公司对排名更高的球员下注，每场比赛投注 1 欧元，最终可能会损失大约 30 欧元。Pinnacle 公司素来以高赔率、高返还、高限额在而业界闻名，由图可得，即使投资收益率均为负，Pinnacle 公司在三种投注策略下的表现都好于 Bet365 公司。它的名声名不虚传，这种情况下 3%的投资损失是非常低的。相反，Bet365 公司的佣金似乎要高得多。

从投资回报率来看，随机投注策略在 Bet365 表现最佳，在 Pinnacle 表现最差。一种可能的解释是，Bet365 的目标客户是那些没有经验的玩家。他们可能已经注意到，他们的玩家往往会在排名更好的球员身上押下过多的赌注，所以为了降低风险，他们会降低排名更好的球员获胜的赔率。

现在设想一下，如果玩家提前知道每场比赛的结果。那他们总是会在比赛获胜者的身上下注。在这种情况下，可以期望的投资回报率是多少？

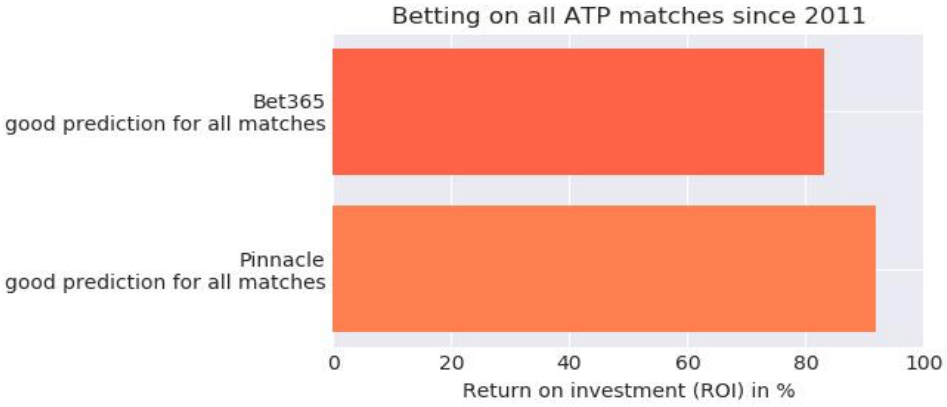

图 5.5 对比赛获胜者下注的投资收益率

由图 5.5 可以看出，假设已知每场比赛获胜者，对所有比赛的获胜者进行投注，

Bet365 和 Pinnacle 投资收益率均超过了 80%，表现更好的 Pinnacle 投资收益率甚至超过了 90%。下面的研究将只关注 Pinnacle 赔率数据。

（2）对选定的比赛下注

与其对所有的比赛进行投注，玩家不如专注于那些自己把握更大的比赛。假设玩家提前知道每场比赛的赢家，选择其中 X%的比赛进行投注，投资收益率如图 5.6 所示。

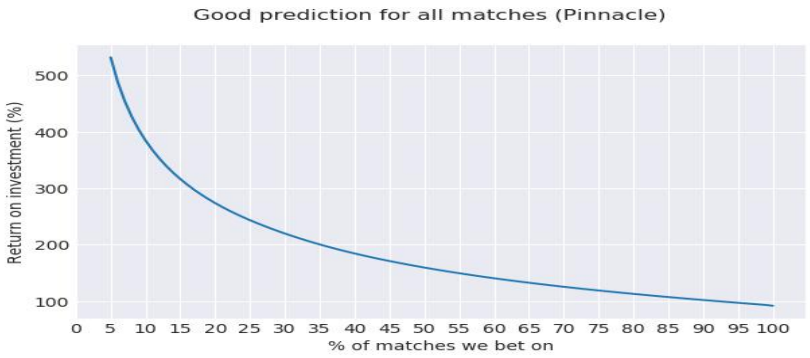

图 5.6 对 X%比赛获胜者下注的投资收益率

从图 5.6 可以看到，如果玩家对 100%的比赛都下注，而且每次都是对获胜者下注，那么可以再次发现投资收益率高于 90%。但如果只选择一部分比赛进行投注，那么投资收益率就会高得多。选定一部分比赛下注通常是一个很好的策略。通过对所有比赛进行投注很难获得良好的投资收益率，但对一部分比赛投注是可行的。唯一的限制是选择下注的比赛的百分比。如果它太小，玩家所押注的比赛数量就会很少，并且投资收益率具有很高的可变性。

5.4.2 提出策略

本节使用XGBoost模型预测每场比赛两种可能结果的概率(球员A获胜/球员B获胜)。通过使用 XGBoost 预测的概率除以隐含概率(赔率的倒数)来衡量投注的可信度，然后选择那些可信度高的比赛进行投注。举个例子：兹维列夫对纳达尔的比赛，博彩公司对兹维列夫获胜开出的赔率为 2.5，这个赔率告诉我们这种情况不太可能发生，兹维列夫是不被博彩公司看好的一方，隐含概率仅为 1/2.5 = 0.4。但 XGBoost 模型预测兹维列夫获胜的概率是 0.8，这可能是因为对纳达尔下注的人过

多，毕竟他是历史上最伟大的球员之一，更受玩家的欢迎和支持。但我们不会被博彩公司愚弄。通过计算可信度达到了 0.8/0.4 = 2，这是非常高的，所以可以赌这场比赛的获胜者是兹维列夫。

（1）预测的准确率

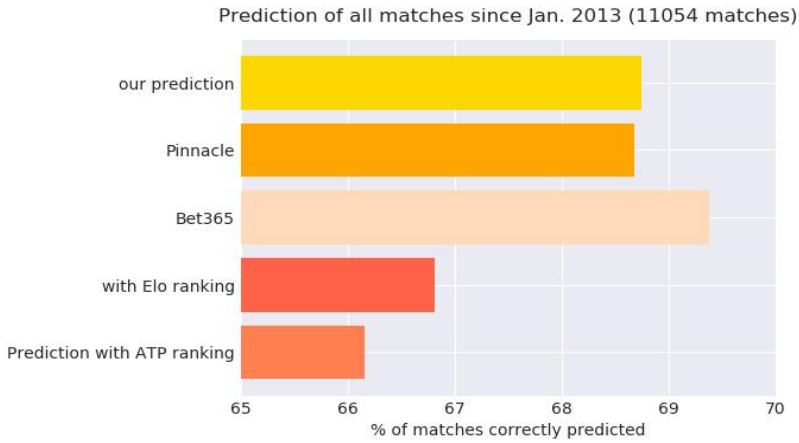

图 5.7 预测准确率

由图 5.7 看出，通过使用 XGBoost 模型，68.7%的比赛被正确预测。虽然预测的准确率并没有非常的高，但预测比赛胜负不是最终的目标，我们想要实现的是最高的投资回报率。如果预测一般是错的，但对于那些高赔率的比赛，全都预测准确，这也是可以的。实际上，模型中所有超参数的选择都是为了最大化投资回报率，而不是为了提高预测精度。

（2）最终投资回报率

根据上述模型的预测及可信度的计算，接下来验证，如果玩家在最有把握的 X%比赛上下注会发生什么。

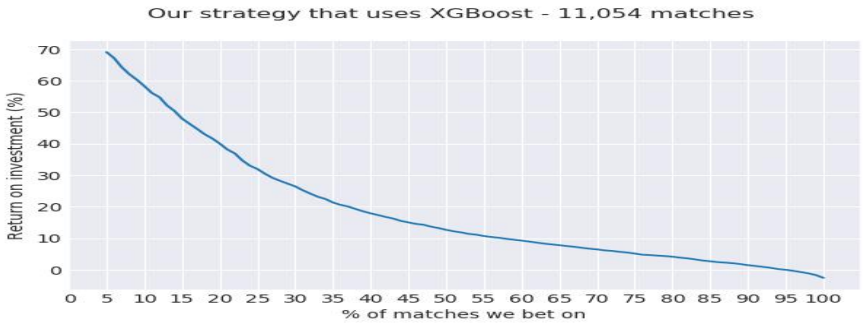

图 5.8 使用 XGBoost 模型的投资回报率

如果使用该策略投注所有的比赛，玩家就会输钱。但如果玩家把赌注押在 5%的比赛上，他们可以获得 70%的投资回报率。

（3）投资回报率波动性研究

与其他运动相比，网球比赛的次数较少(约 2200 场/年)。玩家们不想等待太久来保证他们的投资回报率。假设他们愿意等待连续 117 场比赛(一个大满贯有 117 场比赛，在两周内进行)，并对这 117 场比赛中的一部分进行下注。根据图 5.8 的曲线，对 10%的比赛上投注，平均投资回报率为 58%。玩家决定将赌注押在 2013 年至 2018 年具有最高可信度的 10%的比赛上，大约有 1100 场比赛。首先必须检查这些比赛在 2013 年到 2018 年之间的分布是否良好。

在下面的图中，每个点对应 117 场连续的比赛。如果将该模型应用于 11000 场比赛时，大约有 100 个点。理想的情况是所有点在 y 轴上对应的值约为 11，这意味着模型告诉我们在每 117 场连续比赛中押注 11 或 12 场比赛。

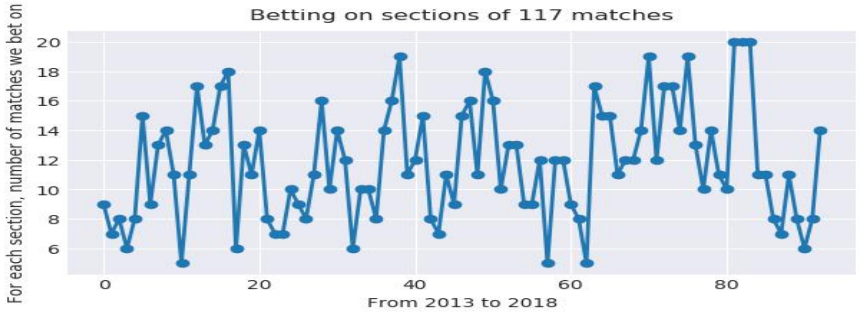

图 5.9 所选比赛分布

由图 5.9，玩家把握最高的的比赛在所选时间内分布良好。这表明博彩市场有一定的稳定性。接下来观察 2013 年至 2018 年投注连续 117 场比赛的投资回报率。

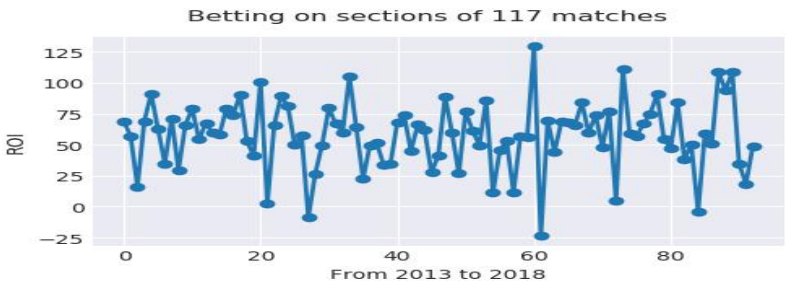

图 5.10 投注 117 场连续比赛的投资回报率

由图 5.10，虽然投资回报率波动很大，但过程似乎相当稳定。综上所述，一个在 2013 年获得良好投资回报率的投注策略在 2018 年仍然可以获得良好的投资回报率，这是相当令人放心的。但这种波动性将迫使玩家等待更多的时间，即 117 场比赛才能保证 58%的投资回报率。现在对可信度最高的 35%的比赛(约 3850 场比赛)进行投注。在这种情况下，研究期间的平均投资回报率为 20%。

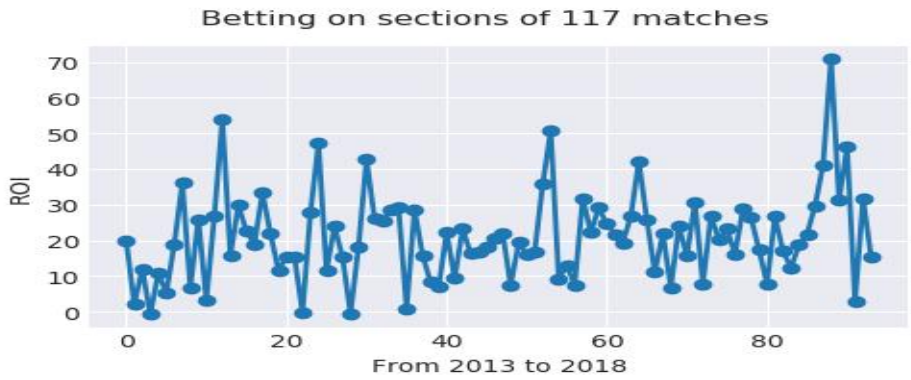

图 5.11 投资回报率

由图 5.11，通过对 35%的比赛中下注，每次连续投注 117 场比赛，玩家几乎从来没有输过钱。

5.4.3 结论

令人满意的结果是，博彩市场似乎相当稳定，这是对网球比赛进行投注的一个重要条件。如果选择合适的投注策略，网球博彩似乎也能带来稳定的收入。通过对 35%的比赛进行下注（在 Pinnacle 上），上述方法可以带来 20%的投资回报率。但仍然有很多改进空间，有很多的相关理论需要探索，以期获得更高的投资收益率。

第六章 总结与展望

6.1 总结

本文主要针对网球比赛赛果预测以及网球比赛的影响因素进行研究。整体来说，主要内容可以分为三个部分：使用因子分析法的网球比赛影响因素分析、结合马尔可夫模型与排名系统的网球比赛预测、使用机器学习模型对网球比赛的预测以及网球投注。

首先进行的是网球比赛影响因素分析。根据一百位球员的职业单打比赛指标数据作为样本数据，根据评价指标体系构建原则和因子分析法，提取出了影响球员表现的三个公因子，分别为发球局表现因子、接发球局表现因子以及发球稳定性因子。依据三个公因子对世界排名前十六的球员进行评分，依据球员得分将这十六位球员按照个人实力划分为三档并对球员进行评价。

然后使用马尔可夫模型根据两名选手在各自发球局赢得一分的概率，递归地计算出选手赢得局、决胜局、盘和比赛胜利的概率。对现有的排名系统进行改进，使用该排名系统来估计球员在发球时赢得一分的概率，并将这些概率输入到一个马尔可夫模型中，以获得球员赢得比赛的概率。还将网球比赛不同类型的场地影响以及发球接发球环节的内在差异纳入我们的模型，对排名系统及模型作出改进，提高预测的正确率。使用改进后的模型对 2016-2019 年的比赛进行预测，预测的准确率分别达到 69.704%、68.377%、68.059%、69.146%。

接下来，基于 2000 年到 2019 年的 ATP 比赛的所有数据，使用机器学习中的支持向量机模型、逻辑回归模型、以及多层感知机模型对网球比赛胜负进行预测，其中表现最好的模型是支持向量机模型，它在测试集上的预测正确率达到 73%。通过对各个特征进行消融研究，得出决定比赛结果的最重要的特征是球员之间的历史交手记录。最后进行网球投注。使用 XGboost 模型和赔率数据对所选比赛进行投注，通过对 35%的比赛下注，平均投资回报率达到了 20%。

6.2 未来展望

首先，关于马尔可夫模型与排名系统的结合，在未来的研究中，可以尝试结

合 Elo 和 Glicko2 排名系统来创建一个新的排名系统。另一种可能性是以某种方式量化球员的发展势头并根据球员发展势头修改 Elo 排名系统中的参数 K 。还可以采用不同的方法来衡量网球场地因素的影响。这部分本文提出了两个应用领域，分别是指导球员训练和比赛时长预测，未来模型在每一个领域的应用都可以单独进行深入研究。

在机器学习模型中，仍然有很多的网球比赛的特征没被纳入，这些特征能提高模型的准确性。例如：球员之前在类似环境（场地、天气）中参加的比赛次数、比赛的重要性程度、球员在这场比赛前面几天参加的比赛次数等等。丰富网球比赛的研究特征对于使用机器学习模型预测网球比赛胜负具有重要的意义。未来的研究中，需要收集更多相关特征的数据，用于机器学习模型的构建。

参考文献

[1]杨麟,赵赟.路径依赖理论视角下的中国女子网球职业化改革[J].北京体育大学学报,2012,35(05):42-45.DOI:10.19582/j.cnki.11-3785/g8.2012.05.009.

[2]陈正,唐小林,周荣,杨志康,余丽桥,刘青.对提高我国网球女子双打竞技水平及其实现奥运突破的思考[J].成都体育学院学报,2004(02):51-54.

[3]蒋宏伟.试论中国网球从“专业化”向职业化的转变[J].南京体育学院学报(社会科学版),2011,25(03):1-6.DOI:10.15877/j.cnki.nsic.2011.03.002.

[4]李苏,徐薇薇.我国女子网球运动发展研究[J].体育文化导刊,2011(01):64-66.

[5]吴贻刚.美国体育博彩的发展对我国体育彩票发展的启示[J].体育文化导刊,2003(11):53-54.

[6]黄珺.中国近代网球运动发展的历史回顾[J].体育文化导刊,2002(01):46-47.

[7]吴云.从世界优秀网球运动员的特点看我国竞技网球运动的发展对策[J].广州体育学院学报,2005(05):98-100+27.DOI:10.13830/j.cnki.cn44-1129/g8.2005.05.029.

[8]刘青,王良佐,唐小林,余丽乔,张琪,杨志康.提高我国优秀网球女子双打运动员竞技水平的研究[J].中国体育科技,2005(02):75-78+84.DOI:10.16470/j.csst.2005.02.023.

[9]孙晋芳.锐意改革 不断进取,努力实现中国网球更大突破[J].北京体育大学学报,2009,32(02):1-7.DOI:10.19582/j.cnki.11-3785/g8.2009.02.001.

[10]郭立亚,袁毅,关晓燕,陈马强.世界顶级网球男子单打比赛制胜技术因素分析[J].北京体育大学学报,2010,33(02):122-124.DOI:10.19582/j.cnki.11-3785/g8.2010.02.033.

[11]祁航. ATP 男子单打比赛制胜因素判别分析及结果预测[D].中国地质大学（北京）,2014.

[12]何文盛,张力为,张连成.世界前 3 名男子网球运动员比赛制胜因素技术分析[J].武汉体育学院学报,2011,45(09):67-73.DOI:10.15930/j.cnki.wtxb.2011.09.016.

[13]蒋启飞,郑贺.男子网球单打技术与综合实力回归预测模型构建分析[J].吉林体育学院学报,2015,31(02):39-43.DOI:10.13720/j.cnki.22-1286.2015.02.009.

[14]张银满.世界优秀男子网球单打选手硬地赛制胜因素[J].北京体育大学学报,2009,32(10):135-137.DOI:10.19582/j.cnki.11-3785/g8.2009.10.040.

[15]杨志敏.男子网球单打比赛成绩预测方程建立[J].北京体育大学学报,2010,33(04):143-145.DOI:10.19582/j.cnki.11-3785/g8.2010.04.039.

[16]罗伟权,张磊.职业网球运动员制胜因素模型构建研究[J].广州体育学院学报,2020,40(03):78-81.DOI:10.13830/j.cnki.cn44-1129/g8.2020.03.021.

[17]岳斌. 优秀男单网球选手硬地比赛技术统计指标分析与胜负回归方程的建立[D].北京体育大学,2013.

[18]孟凡明,黄文敏.基于决策树法的女子网球运动员致胜因素量化分析[J].吉林体育学院学

报,2019,35(05):43-47.DOI:10.13720/j.cnki.22-1286.2019.05.007.

[19]Barnett T . DEVELOPING A TENNIS MODEL THAT REFLECTS OUTCOMES OF TENNIS MATCHES. Australian and NZ Industrial and Applied Mathematics, 2005.

[20]NEWTON PK, KELLER JB. Probability of winning at tennis I. Theory and data[J]. Studies in Applied Mathematics,2005,114(3):241-269.

[21]NEWTON PK, ASLAM K. Monte Carlo tennis[J]. SIAM Review,2006,48(4):722-742.

[22]Newton P K, Aslam K. Monte Carlo Tennis: A Stochastic Markov Chain Model[J]. Journal of Quantitative Analysis in Sports, 2009, 5(3): 1-44.

[23]Barnett T , Clarke S R . Combining player statistics to predict outcomes of tennis matches[J]. Ima Journal of Management Mathematics, 2005, 16(2):113-120.

[24]Somboonphokkaphan A, Phimoltares S, Lursinsap C. Tennis winner prediction based on time-series history with neural modeling[C]//Proceedings of the International MultiConference of Engineers and Computer Scientists. 2009, 1: 18-20.

[25]Glickman M E. Parameter estimation in large dynamic paired comparison experiments[J]. Journal of the Royal Statistical Society: Series C (Applied Statistics), 1999, 48(3): 377-394.

[26]Mchale I , Morton A . A Bradley-Terry type model for forecasting tennis match results[J]. International Journal of Forecasting, 2011, 27(2):619-630.

[27]Sipko M, Knottenbelt W. Machine learning for the prediction of professional tennis matches[J]. MEng computing-final year project, Imperial College London, 2015.

[28]Cui Y, Gómez M Á, Gonçalves B, et al. Performance profiles of professional female tennis players in grand slams[J]. PLOS ONE, 2018, 13(7): 1-18.

[29]王庆海.网球知识介绍[J].北京体育,1982(05):34.

[30]唐琪. 基于多层感知机与客户聚类的客户流失预测算法研究[D]. 广西师范大学,2020.

[31]李航．统计学习方法［M］．清华大学出版社. 2012.

[32]方弦.博彩公司的秘密:赔率是这样炼成的[J].数学教学通讯,2012(13):6-7.

[33]]段国伟. 世界著名男子网球运动员单打比赛制胜因素分析[D].中国矿业大学,2020.DOI:10.27623/d.cnki.gzkyu.2020.002128.

[34]Knottenbelt W J , Spanias D , Madurska A M . A common-opponent stochastic model for predicting the outcome of professional tennis matches[J]. Computers & Mathematics with Applications, 2012, 64(12):3820-3827.

[35]https://baike.baidu.com/view/6016468.html

[36]https://www.biaodianfu.com/elo-glicko-trueskill.html

[37]李松璞. 日本选手锦织圭网球比赛技战术风格对我国男子选手的启示[D].郑州大学,2016.

[38]Bester D W , Maltitz M . Introducing Momentum to the Elo rating System.

[39]J. Sackmann, “Tennis atp.” https://github.com/JeffSackmann/tennis atp, 2018.

[40]Clarke S R, Dyte D. Using official ratings to simulate major tennis tournaments[J]. International transactions in operational research, 2000, 7(6): 585-594.

附录

A.1 B 取值

B	1	2	3	4	5	6
1	1	3	0	3	0	0
2	3	3	1	3	0	0
3	3	4	0	2	1	0
4	6	2	2	4	0	0
5	12	3	1	3	1	0
6	3	4	0	2	2	0
7	4	2	3	4	0	0
8	24	3	2	3	1	0
9	24	4	1	2	2	0
10	4	5	0	1	3	0
11	5	1	4	5	0	0
12	40	2	3	4	1	0
13	60	3	2	3	2	0
14	20	4	1	2	3	0
15	1	5	0	1	4	1
16	1	0	5	5	0	1
17	25	1	4	4	1	1
18	100	2	3	3	2	1
19	100	3	2	2	3	1
20	25	4	1	1	4	1
21	1	5	0	0	5	1

A.2 A 取值

A	1	2	3	4	5	6
1	1	3	0	4	0	0
2	3	3	1	4	0	0
3	4	4	0	3	1	0
4	6	3	2	4	0	0
5	16	4	1	3	1	0
6	6	5	0	2	2	0
7	10	2	3	5	0	0
8	40	3	2	4	1	0
9	30	4	1	3	2	0
10	4	5	0	2	3	0
11	5	1	4	6	0	0
12	50	2	3	5	1	0
13	100	3	2	4	2	0
14	50	4	1	3	3	0
15	5	5	0	2	4	0
16	1	1	5	6	0	0
17	30	2	4	5	1	0
18	150	3	3	4	2	0
19	200	4	2	3	3	0
20	75	5	1	2	4	0
21	6	6	0	1	5	0
22	1	0	6	6	0	1
23	36	1	5	5	1	1
24	225	2	4	4	2	1
25	400	3	3	3	3	1
26	225	4	2	2	4	1
27	36	5	1	1	5	1
28	1	6	0	0	6	1