DOI:10.13546/j.cnki.tjyjc.2019.05.019

基于神经网络的2020年奥运会奖牌成绩预测

王 方

（重庆大学 体育学院，重庆 400044）

摘 要：文章探讨奥运会奖牌成绩的影响因素，定量预测和研究美国、中国、英国、俄罗斯、德国、法国和日本等七个竞技体育强国人均GDP以及2020年东京奥运会的奖牌成绩。采用神经网络非线性方法拟合和预测了人均GDP数据。根据Bernard和Busse提出的预测模型对七个竞技体育强国在2020年东京奥运会可能获得的奖牌数目进行了预测。

关键词：神经网络；回归分析；奥运会；奖牌预测；人均GDP

中图分类号：G80-05 **文献标识码**：A **文章编号**：1002-6487(2019)05-0089-03

0 引言

回顾历史，不难发现各个国家夺得奥运奖牌的数目相去甚远。自1894年以来，美国共获得1022枚金牌，2523枚奖牌，遥遥领先于其他国家。1984—2016年夏季奥运会各个参赛国家的总奖牌数分布情况：56%~68%的国家从来没获得过任何奖牌，18%~26%的国家仅获得了1~5枚奖牌，获得6~10枚、11~15枚、16~20枚和21~25枚奖牌的国家分别占3%~9%、1%~4%、1%~4%和0%~2%，获得奖牌数目超过25枚的国家占4%~6%。在刚刚过去的里约奥运会上，获得奖牌总数超过20枚的国家仅有11个，分别为美国(121)、中国(70)、英国(67)、俄罗斯(56)、德国(42)、法国(42)、日本(41)、澳大利亚(29)、意大利(28)、加拿大(22)、韩国(21)，其中加拿大和韩国的奖牌总数未超过25枚，澳大利亚和意大利的奖牌总数在25~30枚之间，其余7个国家的奖牌总数都在40枚以上。这几个国家的奖牌之和已经超过了里约奥运会奖牌总数目974的55%，充分表现了奖牌数目分布的不均匀性。因此，大家不禁要问，一个国家的奖牌数目和哪些因素相关呢？进一步地，如何预测2020年第32届东京奥运会的各个国家的奖牌数目？

1 预测模型

王国凡和唐学峰(2009)[1]对国内外关于奥运会奖牌预测的文献进行了系统和详细的综述与分析，将奥运会成绩的研究和预测方法分为时间序列模型[2,3]、基于社会学、经济学和地理学原理的经验模型[4-7]和神经网络模型。其中，基于社会学、经济学和地理学原理的经验模型最受推崇。特别是2004年，Bernard和Busse(2004)[4]提出的柯布-道格拉斯生产函数受到广泛关注和使用。该方法认为一个国家奖牌数的分布依赖于以下社会和经济方面的因素：人口数量、人均GDP、东道主效应等。本文采用的函数形式如下：

$$M^i(t)=a_1^i+a_2^i\log\left(PGDP^i(t)\right)+a_3^i\log\left(POP^i(t)\right)+a_4^iHome^i(t)+a_5^iM^i(t-4) \quad (1)$$

其中，$M^i(t)$表示国家i在t年奥运会上获得奖牌数目与当届奥运会奖牌总数的比值。例如，t=2016年，国家i为美国的奖牌数目121，总奖牌数目为974，则对应的M为121÷974= 12.42%。$M^i(t-4)$表示4年前即上届奥运会时对应的比值。POP为当届奥运会参赛国家的总人口数。PGDP为当届参赛国家的人均国内生产总值。Home表示当届奥运会时，国家i是否为东道主：是东道主，则Home为1；否则为0。a_1^i至a_5^i为该国家对应的参数，可以通过最小二乘法拟合得到。

当然，获得奥运会奖牌的影响因素是非常复杂的，有些学者[3,8]还选择了其他可能的影响因素，例如：与东道主的空间距离、各个国家参赛的女性运动员的数目、各个国家的地理位置和生产要素等。有些研究[9,10]则采用结合上述柯布-道格拉斯生产函数和非线性综合模型、智能化算法如模糊C均值聚类分析理论和基因算法等改进预测精度。另外，张海波和赵焕成(2008)[9]则采用Poisson回归的方法建立了奥运会主办国金牌成绩的预测模型，并根据中国军团2004年雅典奥运会的金牌成绩预测了中国军团2008年北京奥运会的金牌成绩。陈军才等(2012)[10]根据历届奥运会金牌的数据，选取有代表性的数据，用回归分析、平稳序列模型进行趋势预测，对2012年伦敦奥运会中国队金牌数和排名进行了预测。

本文的目的是预测2020年东京奥运会上七个竞技体育强国的奖牌情况。如果选择更多影响因素进行研究，则不仅需要搜集2016年以前的数据，还需要预测这些影响因素未来的2016—2020年之间的数据，所以选择较简单

基金项目：中央高校基本科研业务费专项资金资助项目(106112016CDJXY250001)

作者简介：王 方(1982—)，女，四川南充人，硕士研究生，讲师，研究方向：体育教育训练学。

的公式(1)进行研究和预测。预测模型即公式(1)采用Origin9.0工具拟合。

2 数据来源

本文收集了1984年洛杉矶夏季奥运会至2016年里约夏季奥运会,以下七个竞技体育强国:美国、中国、英国、俄罗斯、德国、法国、日本等的相关数据。其中奖牌数据来源于奥林匹克国际官方网站(https://www.olympic.org/)。联合国贸易与发展组织UNCTAD数据库(http://unctadstat.unctad.org/)可以查询各个国家历年(直至2050年)的人口数据。世界各个国家的人均国内生产总值(PGDP)来自于UNCTAD数据库、美国中央情报局的世界百科全书(http://data.worldbank.org/)和中华人民共和国统计局官方网站(http://www.stats.gov.cn/)等。本文采用的人均国内生产总值是2005年不变价GDP,单位为美元。

由于所有数据库都未提供2015年之后的GDP数据,本文采用了神经网络方法对2015年之后的GDP数据进行了预测。神经网络方法是一种具有高度灵活性的非线性拟合方法,理论上可以拟合任何函数,并且达到非常高的拟合精度。神经网络函数有很多类型,本文采用前馈型神经网络。简要地,以时间为输入层,选择一层隐藏层(隐藏层的神经元数目可以调节),输出层为各个国家的人均GDP,如下:

$$PGDP^{i}(t)=b_1^{(2)}+\sum_{j=1}^{J}\left(\omega_{1,j}^{(2)}\cdot f_1\left(b_j^{(1)}+\omega_{j,1}^{(1)}\cdot t\right)\right) \qquad (2)$$

其中,J代表隐藏层的神经元的个数,与所研究的具体国家相关;f代表隐藏层的转换函数,本文选用双曲正切函数,即f(x)=tanh(x)。其中的ω和b为连接相应层神经元之间的参数,通过非线性拟合得到。对任意国家的PGDP,拟合均方根误差表示为:

$$RMSE=\sqrt{\frac{1}{n}\cdot\sum_{t=1984}^{2014}\left(PGDP^{fit}(t)-PGDP^{real}(t)\right)^2} \qquad (3)$$

神经网络的拟合过程就是不断地优化参数ω和b,使误差RMSE减小到所需求的精度。使用方法为Levenberg-Marquardt算法,具有收敛速度快,执行效率高,通用性好等特点。当RMSE的下降速率小于给定的阈值,或者RMSE小于预先设定的标准时,神经网络函数就完成了一次拟合,即生成了一个神经网络模型。由于神经网络的拟合是一个寻找拟合误差的局域最优的过程,从不同的随机值(即对参数ω和b随机赋予初值)出发,最后会得到不同的拟合模型。一般要进行平行地多次拟合,最后取RMSE最小的一次或几次拟合结果的平均值作为最终的模型。

为了避免出现过度拟合现象,将样本分为训练集、测试集和验证集,分别占90%、5%和5%。仅选择那些三个集和的拟合精度接近的神经网络模型。神经网络拟合即公式(2)则采用Matlab中的神经网络工具包进行非线性拟合。其他的数据统计和分析则通过Excel处理。

3 研究结果与分析

3.1 基于神经网络的PGDP模型

神经网络的输入层为年份,输出层为相应年份的PGDP。确定好输入与输出后,Matlab中的神经网络工具包会对数据进行预处理,以将其转变为神经网络需要的标准化数据,并随机分配90%、5%和5%分别为训练集、测试集和验证集。隐藏层选择为1层,本文测试了不同神经元数目(4、5和6)对拟合精度的影响。通过多次建模发现,拟合一般在几十步迭代后就达到收敛。

图1比较了美国在1984—2014年间的实际人均GDP与神经网络拟合值。经过测试,最终选择的隐藏层神经元数目为5。拟合的均方根误差为177美元,这相比于几万美元的人均GDP可以忽略。如图1所示,拟合值与实际值符合得非常好。2015—2020年的人均GDP为预测值,即美国的人均GDP在未来几年仍然会增长。预测的2016年的美国人均GDP为55655美元,与媒体近日的估测值接近。预测的2020年的美国人均GDP为58059美元。

相似地,图2给出了中国1984—2014年间实际人均GDP与神经网络拟合值的比较。隐藏层神经元数目为5,拟合的均方根误差为55美元。如图2所示,拟合值与实际值符合得非常好。2015—2020年的人均GDP为预测值,即中国的人均GDP在未来几年仍然会增长,且保持较高速度增长。预测的2016年的中国人均GDP为8452美元,与媒体的估测值接近。预测的2020年的中国人均GDP为10204美元。

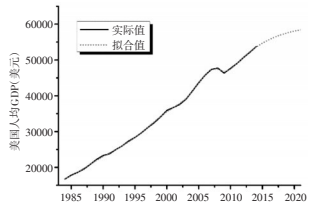

图1 美国实际人均GDP与神经网络拟合值的比较

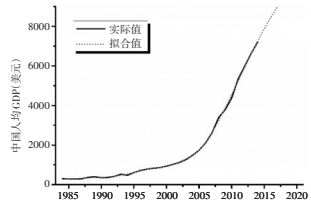

图2 中国实际人均GDP与神经网络拟合值的比较

下页图3至图7分别给出了英国、俄罗斯、德国、法国和日本的历年的人均GDP和神经网络拟合值的比较。由于政治原因,俄罗斯仅有1992年后的人均GDP数据,德国仅有1990后的人均GDP数据。美国和中国的人均GDP随着年份增加而单调递增,因此比较容易拟合和预测。而这5个国家的人均GDP都呈现出复杂的结构:随着年份增加,呈现无规律的震荡,这对于拟合是一个挑战。经过测试,所选择的隐藏层的神经元数目分别为5、5、5、3、5,拟合的均方根误差分别为802、171、987、933、1942美元。对于法国人均GDP,采用了两层隐藏层,每层神经元数目为3。由于这些国家高度发达,人均GDP较高,因此,如图3至图7所示,拟合值与实际值吻合,体现了神经网络方法的强大拟合和预测能力。

3.2 2020年奥运会奖牌预测结果

根据上述分析,利用1984—2016年间的相应数据,可以根据公式(1)对这7个竞技体育强国进行回归分析,结果见表1。其中MAE指平均误差,RMSE为均方根误差。根据表1的结果,以及预测的2020年的人均GDP和人口数据,可对2020年各个国家的奖牌数进行预测,见表1最后一

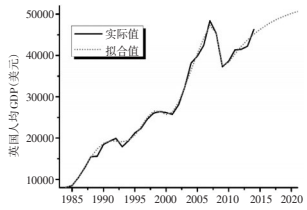

图3 英国实际人均GDP与神经网络拟合值的比较

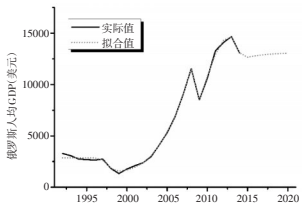

图4 俄罗斯实际人均GDP与神经网络拟合值的比较

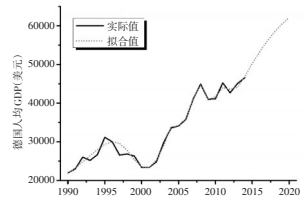

图5 德国实际人均GDP与神经网络拟合值的比较

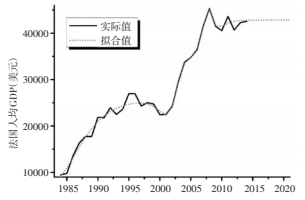

图6 法国实际人均GDP与神经网络拟合值的比较

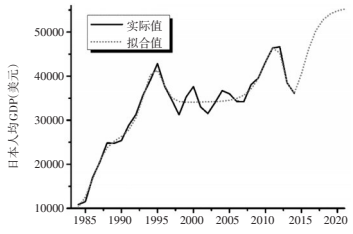

图7 日本实际人均GDP与神经网络拟合值的比较

列。预计2020年东京奥运会上,美国将获得111枚奖牌,比里约奥运会有所减少。一方面是由于4年后俄罗斯田径项目的回归,另一方面是由于主办地在东京城市,远离美国。中国将获得81枚奖牌,比2016年里约奥运会有所增加。这是由于今年羽毛球等传统强项失利,而且主办地在南美洲,造成了球员的不适应性。另外,东京在地理位置上非常接近中国,因此气候上的区别也可以忽略不计。日本将获得45枚奖牌,与2016年基本持平。这是由于历史上巴西有非常多的日本移民,使得日本在今年奥运会上获得了突破性的奖牌数目,预计2020年东京奥运会上东道主效应与2016年持平,还有部分奖牌被亚洲和欧洲的体育强国夺走。预测由于俄罗斯田径等项目的强势回归,将获得68枚奖牌。预测英国、德国、法国将分别获得73、37和40枚奖牌。

表1 1984—2016年间历届奥运会奖牌回归统计及2020年奖牌数目预测

国家	a_1	a_2	a_3	a_4	a_5	MAE	RMSE	2020
美国	−30.6299	−0.5728	1.8889	−0.0075	−0.2674	6.2	9.5	111
中国	−1.5276	0.0043	0.0745	0.0304	0.0236	4.2	6.0	81
英国	−10.7066	−0.0382	0.6219	0.0050	−0.0166	3.9	5.0	73
俄罗斯	−0.0010	−0.0143	0.0099	0.0023	0.2529	5.0	7.6	68
德国	81.9765	−0.1000	−4.4443	0.0000	0.8427	5.8	7.0	37
法国	−0.3019	0.0068	0.0158	0.0000	−0.3419	2.5	3.5	40
日本	−7.9336	−0.0304	0.4448	0.0203	−0.0874	5.0	6.2	39

3.3 分析与讨论

如前所述,神经网络方法对于各个国家人均GDP具有强大的拟合能力,能够非常好地重现GDP随时间震荡的情况。所以,神经网络方法可以被广泛应用于体育研究的许多方面。预测的2020年的人均GDP数值非常准确,保证了后续对奖牌数预测的准确性。本文预测的人均GDP也为今后的其它研究提供了可靠的数据来源。

在奖牌预测方面,本文采用了线性回归的方法。这是由于样本量非常少,仅有1984—2016年共9届奥运会的数据,而神经网络方法一般对较大样本的问题具有更好的拟合和预测能力。另外,由于数据来源受限等问题,本文选择的影响因素较少,使得预测模型不够准确。相信随着样本数目的增加以及有更多可靠的数据来源,未来对于奥运会奖牌成绩的预测将会更加准确。

另外,预测模型是一种数学工具,它可以根据过去的表现与可能的影响因素,对未来的表现进行不含感情因素的预测。由于其影响因素复杂,且经常伴有突发事件,因此难以准确和定量的预测。2016年奥运会上,中国代表团的诸多传统强项如羽毛球失利。但是,也要看到如林丹老去、女单青黄不接、游泳冠军宁泽涛的离队等问题将严重影响四年之后的奥运会奖牌榜。因此,应当做好万全准备,积极迎接2020开在我们海上邻国的东京奥运会。

4 结论

本文采用神经网络非线性方法拟合得到了2015年后的人均GDP数据,然后根据Bernard和Busse提出的柯布–道格拉斯生产函数对7个体育强国在2020年东京奥运会可能获得的奖牌数目进行了首次预测,并进行了分析探讨。神经网络方法的强大拟合和预测能力使得拟合精度非常高,拟合很好地重现了实际值,特别是当实际值呈现无规律震荡情况时。神经网络方法快速、有效、精确。应当注意,影响GDP和奥运会成绩的影响因素非常复杂,任何模型都有其局限性。

参考文献:

[1]王国凡,唐学峰.奥运会奖牌预测国内、外研究动态及发展趋势[J].中国体育科技,2009, 45(6).

[2]范珣,齐辉.运用趋势直线外推法预测2008年奥运会中国获奖牌数[J].辽宁体育科技, 2007, 29(4).

[3]Johnson D K N, Ali A. Coming to Play or Coming to Win: Participation and Success at the Olympic Games[S].Wellesley College, Mimeo, 2000.

[4]Bernard A B, Busse M R. Who Wins the Olympic Games: Economic Resources and Medals Totals[J].The Review of Economics and Statistics, 2004, 86(1).

[5]Emrich E, Klein M, Pitsch W, et al. On the Determinants of Sporting Success –A Note on the Olympic Games[J].Economics Bulletin, 2012, 32(3).

[6]王国凡, 薛二剑, 唐学峰.大型国际综合性运动会奖牌数预测研究——以北京奥运会为例[J].天津体育学院学报, 2010, 25(1).

[7]陈丹, 赵海燕. 世界竞技体育实力空间自相关分析: 基于1–30届夏季奥运会成绩[J]. 中国体育科技, 2015, 51(5).

[8]Gerard K, Elmer S. Olympic Participation and Performance Since 1896[J].SSRN Electronic Journal,2001.

[9]张海波, 赵焕成. 北京奥运会中国军团金牌数的预测[J]. 统计与决策, 2008,(15).

[10]陈军才, 林海明, 景旻. 伦敦奥运会中国队金牌数和排名的预测[J]. 统计与决策, 2012,(12).

(责任编辑/浩 天)