**Your Paper's Title**

**Summary**

Keyword**：**

**目录Contents**

**1.引言(Introduce)**

**1.1问题背景(Problem Background)**

近年来，奥运会在全球内的关注程度不断提高，一次奥运会的举办对于奥运圣地和世界各国将会在政治、经济、社会文化等多方面产生深远的影响。对于奥运圣地为保障奥运会的顺利举办将调整体育政策、预算分配及社会资源的配置，使得交通、住宿和体育设施等基础设施得到改善，因此大量的施工、服务及管理岗位会在准备和举办期间创造出众多就业机会，这些完备的基础设施在赛事结束后也可提升城市水平，并能激励更多人参与体育活动，促进全民健身意识，从而提高整体健康水平提高人民的生活品质，同时举办奥运会通常能吸引大量游客，推动当地旅游业发展，并吸引国内外投资，主办国通过成功举办奥运会能提升其国际地位和影响力，展示其政治稳定和经济发展。对于世界各国而言，运动员的成就带来的即时关注，能够增强国民的民族自豪感，提升国家形象，同时促进不同文化之间的交流与理解，有助于增进国际友谊。

粉丝们不仅将目光聚焦于夏季奥运会的体育赛事上，每个国家的总奖牌榜与金牌榜的排名也可谓是备受关注，当位于榜单前列的众多国家备受重视的同时，阿尔巴尼亚、佛得角、多米尼克和圣卢西亚等小国也在众多国家中脱颖而出拿下了属于本国历史上第一枚奥运奖牌。由此可见我们的奥运赛场是充满着无限可能的，而这些奖牌背后又隐藏着这样的规律仍有待揭晓，人们从各方面刨析奥运会获奖原因，使得奥运运动成绩的预测和奥运奖牌榜的预测是近年来的热门。

奥运会作为全球体育盛典，吸引了世界各国的参与。其所获运动成绩及奖牌数不仅是一个民族，一个国家竞技水平的反映，更是国家意志力的表现。因此，各国都十分重视奥运会的成绩预测。此项研究不断升温，并在短时间内就有不同的学科领域提出多种预测模型方法来帮助预测,其理论方法和预测模型在迅速发展和完善。

预测科学由来已久，现代预测学科学主要研究无所不在的不确定性，目的是将随机性以及无知的程度降到最低。以达到减少人们生活方方面面因为不确定性导致错误决策所产生的风险。近四十多年来，国际上对预测理论和试验的分析以及探讨发展十分迅速。有些国家更是聚集了当代最具权威的自然科学家和社会科学家加以最先进的技术，从事预测方面的研究，使预测学的原理和方法得以全面开花，并取得了显著的成效。要使我国体育事业健康稳步发展需要进行科学的体育预测，体育预测势在必行。

预测是决策的基础，没有科学的预测，没有对未来情况的研究和把握，就不可能采取正确的决策和行动，体育预测能提高各项工作的预见性。这样才能识时明势，正确预测运动项目的发展趋势和各国的发展动向，采取一系列有效对策。体育预测能促进训练工作科学化，一些国家通过编制专门预测运动成绩的电子计算机程序，能科学预测运动员的成绩水平，从而根据预测结果设定相应的训练计划，训练内容和训练任务，大大提高训练工作的科学性和有效性。体育预测能为正确决策提供科学依据。教练员只有正确的预测了对方的战略战术意图，科学的分析了比赛中可能发现的有利或不利因素，才能做到把握全局，从而掌握比赛的主动权。在制订体育发展规划，编制各项工作计划时，更需要预测提供的各种资料和信息，为决策提供科学依据。

**1.2Restatement of the Problem(问题重述)**

考虑到问题陈述中确定的背景信息和限制条件，我们需要解决以下问题：

为每个国家（金牌和总奖牌数）制定一个奖牌计数模型，对2028年美国洛杉矶夏季奥运会奖牌榜的预测，说出某些国家会进步或者退步及其原因。

预测在下一届奥运会上有多少国家会获得第一枚奖牌及其概率，各国运动强项是哪些项目以及东道主的选取是否会造成各国奖牌数量波动。

由于教练对其执教队伍的执教队伍又主动选择权，则会出现“伟大教练”前往其他国家队进行执教而导致各国代表队在奥运赛场上的表现出现变化，结合有关数据分析是否存在“伟大教练”效应。

**1.3 Literature Review**

Müller, M., Gollner, B., & Tschernutter, C.在2020于《International Journal of Sports Science & Coaching》发表"Predicting Olympic Success: A Machine Learning Approach"以国家历史成绩、运动员的个人记录和背景信息、体育项目的竞争激烈程度、投入的资源（如训练设施、资金支持）为重要特征，通过线性回归、决策树、随机森林、支持向量机（SVM）、神经网络等模型进行交叉检验对如何准确预测各国在奥林匹克运动会中的奖牌数量和类别（如金牌、银牌、铜牌）、哪些因素能够影响运动员和国家的成功率等问题进行讨论。

Julia Bredtmann、Carsten J. Crede、Sebastian Otten在《Significance》2016 年第 3 期中发表“Olympic medals: does the past predict the future?” 以过去四届奥运会的总奖牌数、各国的体育参与度、政府对体育的投资和支持为重要特征，通过回归分析、时间序列分析、分类模型、随机森林等通过交叉验证等方法评估模型研究是否不同国家之间资源分配不平等导致了在获得奖牌机会上存在的差异。

Christoph Schlembach、Sascha L. Schmidt、Dominik Schreyer、Linus Wunderlich在《Technological Forecasting and Social Change》2022 年第 175 卷中发表“Forecasting the Olympic medal distribution – A socioeconomic machine learning model” 以国内生产总值 (GDP)、人均收入、人口数量、教育水平为重要特征，通过多种机器学习算法构建预测模型，通过交叉验证和其他统计评估方法（如均方误差、R²值）来评估模型的性能研究国家的经济实力、体育设施的投资、参与体育活动的人口比例等社会经济因素对奖牌分配有显著影响。

**1.4 Our Work(问题分析)**

考虑问题的背景和前人的研究成果，我们针对不同的问题进行了分析：

定义并选取运动员状态和项目数量作为自变量并标准化

加权描述两个自变量的贡献度

建立模型并预测奖牌榜

引入ID3算法以及熵权法的思想对权重进行求解

代入数值求出具体参数值并进行预测

利用权重比以及定义不准确度对模型进行检验

分析附件确定自变量并得到矩阵

利用TOPSIS逼近法处理矩阵得到相对接近度

预测首次拿奖国

对处理过后的矩阵排序并写出前十名

通过比较最终进行预测

**2 Assumptions and Justifications模型假设**

**3 Notations符号说明**

|  |  |
| --- | --- |
| **Symbol** | **Description** |
| G | 金牌数量（有重复） |
| S | 银牌数量（有重复） |
| B | 铜牌数量（有重复） |
| SUM | 总比赛数 |
|  | 奖牌获得率 |
| X | 运动员状态值 |
| E | 项目总数量 |
|  | 奖牌数量 |
| d | 变异指数 |
| e | 信息熵 |
|  | 奖牌变量 |
| P | 标准项目比 |
| K | 最大环境容量 |
| r | 自然增长率 |
| A | 剩余增长空间比 |
| π | 不准确度 |
| d | 距离 |
| C | 相对近似度 |
|  |  |
|  |  |

上表中的部分符号说明：一个项目可有多个运动员参赛，每个运动员也可以参加多个比赛，所以这里的总比赛数SUM不指项目总数，也不是运动员数量，假设N个项目，平均每个项目有n个运动员参赛，SUM = N×n，由于运动员参赛有组队情况，而在奖牌榜中一个冠军只对应一个奖牌，G，S，B则表示这种重复的情况时的奖牌数量，则表示记录在奖牌榜的奖牌数量。

1. **信息熵加权预测模型**

**4.1数据分析以及预测模型的建立**

4.1.1 影响奖牌数量的自变量的讨论

奥运会奖牌数量预测一般不参考历史奖牌总数，所以不考虑历史奖牌的数量，而将其作为函数的因变量进行模型的检验。大多数情况运动员当前的状态就可直接影响最终的各奖牌数量的结果，而如何分析某个国家运动员的表现，我们参考运动员的历史成绩，以某国上一次奥运会的成绩来宏观反映出该国家的运动员状态，从附件可以看出每一次奥运会都有运动员参赛，其中有大部分运动员没用收获奖牌，少部分运动员收获金牌，银牌，铜牌，用该国家获得金牌的项目数量除以总项目数量得到金牌获得率，又由于三种奖牌地位有高低之分，对金牌，银牌和铜牌的获得率加权相加，分别赋予权重3,2,1，得到一个新的变量，我们定义为运动员状态值X，这里我们以美国为例进行研究，其中

取值在0到1之间，X可以直接影响最终奖牌数量。

此外，奥运会总项目越多，整个奥运会分发出的奖牌也就越多，那各个国家可获得奖牌的可能性也就越高，也可从图像看：

图表, 散点图

描述已自动生成

图1 项目数量和美国总奖牌数量关系图

明显奥运会总奖牌的数量也会影响某一国家的奖牌数量，而奥运会项目总数直接决定了奥运会奖牌的总数量，但两者对奖牌数量的贡献度明显不同，项目总数E与X相差过大，应将E标准化，设E=300时，P=1且

于是就可定义奖牌变量：

4.1.2 算法的选择与模型的概述

决策树模型ID3作为一种贪心算法，起到构造决策树的作用。ID3算法起源于概念学习系统（CLS），在每个节点选取还尚未被用来划分的具有最高信息增益的属性作为划分标准，然后继续这个过程，直到生成的决策树能完美分类训练样例。

熵权法作为权重确定方法，以信息熵为基准，通过计算各指标的熵值，比较指标间的差异程度，进而确定各指标的权重。熵值与指标间的差异呈反相关，可以说明该指标在决策中的重要性越高。

本文将两种方法结合，先使用ID3算法中的信息熵与信息增益的算法来得出决策属性的无序程度和各属性对决策属性的影响程度。又借鉴了熵权法中对于熵值的处理方法得出两种能力的权重，这种方法可以结合两种方法的优点，使权重确定更加合理和全面。

我们将运动员状态X划分为三个部分，以美国为例，将状态设为：良好，一般，糟糕三个等级，将奥运会项目总数量E划分为，多，中，少三个等级，将美国所获得金牌数量分为，多，中，少三个等级，算出各个水平比重并运用ID3算法计算出E和X各自对于的信息增益。

关于具体运用熵权法的思想：熵权法对样本指标标准化的过程是将描述样本的值控制在[0,1]，在用其进行熵值的计算，而信息熵算法的范围是，本文利用熵权法对于熵值处理的方法来计算信息熵的权重，在熵权法中要对熵值求变异指数：d=1-e，这里的“1”为描述样本的值标准化后的的最大值，则我们算信息熵的变异指数的时候公式应该为：

其对应的权重k1和k2由熵权法中对变异指数处理得到权重的公式算出：

最后得到美国金牌总数量的式子：

4.1.3 求解模型

因为世界各国发展迅速，所以以2000年以来的奥运会数据来计算得出的结果对如今更具有参考价值，通过对附件的筛选和计算我们得出表格：

表1 美国近年来的运动员状态计算

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Year** | **SUM** | **Gold** | **Silver** | **Bronze** | **X** |
| 2000 | 776 | 130 | 61 | 51 | 0.725515 |
| 2004 | 735 | 117 | 75 | 71 | 0.778231 |
| 2008 | 768 | 127 | 110 | 80 | 0.886719 |
| 2012 | 698 | 145 | 57 | 46 | 0.852436 |
| 2016 | 726 | 139 | 54 | 71 | 0.820937 |
| 2020 | 856 | 113 | 110 | 75 | 0.740654 |
| 2024 | 854 | 131 | 96 | 94 | 0.795082 |

以美国自身为参考划分运动员状态等级，规定运动员状态值在0.75和0.8之间为一般，大于0.8为良好，小于0.75为糟糕。又观察2000年以来奥运会项目数量，以自身为参照，规定项目数量在310和320之间为中，小于310为少，大于320为多得到如下表格，同理，规定金牌数量在37和45之间为中，小于37为少，大于45为多：

表2 决策属性表

|  |  |  |  |
| --- | --- | --- | --- |
| **Year** | **X** | **E** |  |
| 2000 | 糟糕 | 少 | 中 |
| 2004 | 一般 | 少 | 少 |
| 2008 | 良好 | 少 | 少 |
| 2012 | 良好 | 少 | 多 |
| 2016 | 良好 | 少 | 多 |
| 2020 | 糟糕 | 多 | 中 |
| 2024 | 一般 | 多 | 中 |

依据上表可计算得出信息增益以及信息熵的计算结果：

表3 以美国金牌为例，所有各项指标的数据结果

|  |  |
| --- | --- |
| **指标** | **得到的结果** |
| 无条件熵 | 1.5566567074628228 |
| 运动员状态的信息增益和信息熵 | (1.138724635395457, 0.27709636376654934) |
| 项目总量的信息增益和信息熵 | (0.18150882310232208, 1.2343121760596842) |
| k1，k2 | (0.6756010197686305, 0.32439898023136954) |

可得结果：

以2000年美国金牌数量为基准定义奖牌常数使得

则

代入数据即可求出奖牌数量。

同理可算出：

表3 银牌和铜牌所有各项指标的数据结果

|  |  |  |
| --- | --- | --- |
| **指标** | **银牌** | **铜牌** |
| 无条件熵 | 1.556656707 | |
| k1，k2 | 0.45730441975063163, 0.5426955802493684 | 0.826933753115614, 0.17306624688438585 |
| Q | 27.4 | 41.4 |

4.1.4 对2028年奖牌榜的预测。

（11），（12）（13）式中两个主要的自变量为运动员状态值X和标准项目值P，我们可以通过对项目总数E与年份进行二维图像分析反映P值：

图表, 折线图

描述已自动生成

图2 项目总量与年份关系图

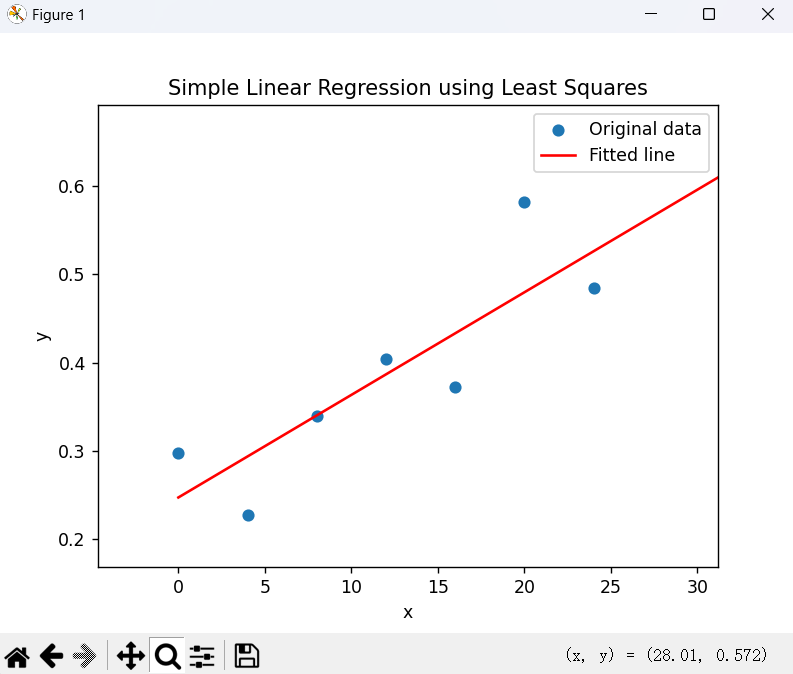
如图所示以2000年为起点，前段图像近似指数图像，后半部分开始有所下滑但较为平缓，这类图像可以用Logistic人口模型近似描述，Logistic人口模型的表达式为：

观察图像设K=350，E(0)=300较为合理，则A=0.17，取值(2016,306)代入(14)式可算出r=0.0105。即

将t=2028代入(15)式得，E=344，则P=1.1467。

接下来取2024年奖牌榜前十名作为样本计算出各国的运动员状态值

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 年份  国家 | 2000 | 2004 | 2008 | 2012 | 2016 | 2020 | 2024 |
| 美国 | 0.725515 | 0.778231 | 0.886719 | 0.852436 | 0.820937 | 0.740654 | 0.795082 |
| 中国 | 0.380368 | 0.389068 | 0.473236 | 0.524345 | 0.427562 | 0.460317 | 0.543296 |
| 日本 | 0.23416 | 0.374396 | 0.246696 | 0.35589 | 0.254587 | 0.39136 | 0.279264 |
| 澳大利亚 | 0.472081 | 0.552413 | 0.446208 | 0.363813 | 0.312741 | 0.362069 | 0.355932 |
| 法国 | 0.297872 | 0.227766 | 0.339326 | 0.403756 | 0.373047 | 0.581784 | 0.484395 |
| 荷兰 | 0.552901 | 0.509363 | 0.53169 | 0.630631 | 0.273556 | 0.409207 | 0.616927 |
| 英国 | 0.284337 | 0.322222 | 0.402878 | 0.368421 | 0.686192 | 0.485401 | 0.515947 |
| 韩国 | 0.342618 | 0.313953 | 0.519757 | 0.352564 | 0.209125 | 0.215976 | 0.45968 |
| 意大利 | 0.269147 | 0.407173 | 0.156863 | 0.319372 | 0.320802 | 0.26145 | 0.278689 |
| 德国 | 0.358025 | 0.466216 | 0.342342 | 0.413725 | 0.559701 | 0.249581 | 0.307573 |

利用最小二乘法可预测出2028年各国的运动员状态值并作出图像，以法国为例：

上图横坐标为年份-2000，纵坐标为运动员状态值。

经过计算之后可以得到新的奖牌榜：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **2024年排名** | **2028年预测排名** |  |  |  |
| 美国 | 美国 | 42 | 27 | 37 |
| 中国 | 英国 | 36 | 24 | 31 |
| 日本 | 法国 | 34 | 24 | 29 |
| 澳大利亚 | 中国 | 33 | 23 | 27 |
| 法国 | 荷兰 | 31 | 22 | 25 |
| 荷兰 | 德国 | 27 | 21 | 20 |
| 英国 | 日本 | 27 | 21 | 20 |
| 韩国 | 澳大利亚 | 26 | 21 | 19 |
| 意大利 | 意大利 | 25 | 20 | 18 |
| 德国 | 韩国 | 22 | 19 | 15 |

通过新奖牌榜与旧奖牌榜进行比较，可知这十个国家中中国，日本，韩国，澳大利亚会在2028年奥运会退步，英国，法国，荷兰，德国将会进步。

4.2模型检验

4.2.1 可靠性检验

根据无条件熵的含义，该模型选取的决策属性数量为3，故无条件熵值根据公式最大值为，根据无条件熵的数据与最大值作比较，比值约为98.2%，二者的充分接近，这说明建立的模型随机抽样结果成功，能够较好的反映实际情况，包含更多信息量。而条件属性的信息增益是对决策属性的影响程度的反映，条件属性的信息熵是其对决策属性的无关程度的反映。观察发现，运动员状态值与项目总数对决策属性的贡献都是明显的，因为它们的信息增益值相对较高，这意味着它们是重要的分类特征。

以2000年来的中国情况为样本对奖牌数量模型进行检验：通过计算得出下表：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Year** | **X** |  |  |  | **E** |
| 2000 | 0.380368 | 28 | 16 | 14 | 300 |
| 2004 | 0.389068 | 32 | 17 | 14 | 301 |
| 2008 | 0.473236 | 48 | 22 | 30 | 302 |
| 2012 | 0.524345 | 39 | 31 | 22 | 302 |
| 2016 | 0.427562 | 26 | 18 | 26 | 306 |
| 2020 | 0.460317 | 38 | 32 | 19 | 339 |
| 2024 | 0.543296 | 40 | 27 | 24 | 329 |

将X代入模型得出结果：

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 26 | 20 | 20 |
| 27 | 20 | 21 |
| 30 | 21 | 23 |
| 31 | 22 | 25 |
| 29 | 21 | 22 |
| 29 | 23 | 24 |
| 33 | 23 | 26 |

对比两表可看出，模型对此抽样的预测有一定参考价值，但仍然有一些不准确的地方，我们取预测值和实际值的差与实际值的商来反映结果的不准确度，在取所有不准确度的平均值得到模型的不准确度，设为π，则

模型的误差较小，有着不错的参考价值，但与实际情况有一点小偏差，这是不考虑题目所给附件之外数据的情况，当获得其他样本数据时便可完善模型，做到误差更小。