文章编号:1672 -058X(2011)03 -0318 -04

体育比赛结果预测模型的尝试性研究

朱 文 富
(重庆工商大学 体育学院,重庆 400067)

摘 要:利用计算机数据决策分析方法,针对以个体成绩为主进行名次排序的体育比赛,总结出了分析比赛成绩的建模方法和预测比赛结果的建模方法,以此来分析和确定对每位参赛运动员比赛的成绩有影响的重要因素,并尝试性地预测比赛的结果。

关键词:KDD; 比赛; 预测; 建模
中图分类号:G807 文献标志码: A

如今体育比赛已经成为人们生活中的重要内容,从女排的五连冠到男足的韩日世界杯之旅,再到中国北京奥运会的成功举办,中国体育事业的突飞猛进已经向世界证明东方巨龙巍然屹立在体育赛场上,五星红旗高高飘扬在竞技之颠。然而人们关注体育比赛的重要方面在于它的结果,比赛结果已经成为体育比赛的焦点和集中点。如何对体育比赛进行合理的、科学的预测是当今广大体育科研工作者广泛关注的话题。在这个社会信息化高速发展的时代,KDD(Knowledge Discovery in Data-base) 这项原本不为大家熟悉的技术现如今已受到极其广泛的关注,同时也开始了进一步的研究,此外,此项技术已经在很多相关的领域得到了运用并且相当成功[1-2]。从这一方面可以看出,KDD 使用的目的不是单方面的,它不仅要在实践中得出数据,并在这些数据中通过进一步的分析找出一部分具有重要价值的决策支持信息。所以,可以认为 KDD 它不是研究某种具体的方法,而是主要着重于系统的实用性,是根据每一个用户需求以及该研究领域的特点,利用现在已经掌握的技术,在计算资源相当有限的情况下,从实践中得到的众多数据中发现一些可运用到研究中的各种有用的信息。在此就是针对此类非对垒式的,以每一位参加比赛的以个体的成绩为主要研究对象来进行排定名次的体育竞赛和娱乐比赛,在这种情况下应用 KDD 技术建模,制定出一个该类比赛的情况分析并且对最后比赛成绩进行预测。其意义就是要验证并且发现对该类别比赛成绩有影响的各种因素,同时对最后比赛的结果进行科学的预测。从而在对个体如何最有效的提高比赛成绩,安排其参加何种比赛较适合方面起着辅助决策作用。

1 构建模型的任务及特征

在体育比赛中属于非对垒式比赛或项目占了很大的比重,非对垒式比赛(即以个体成绩为主的名次排序比赛) 。对于该类比赛,平时的训练以及比赛前的准备应该注意对成绩有意义的影响因素,要注意哪些问题才有利于提高成绩; 从另一方面来说,对于每一个个体如何科学的安排和合理的选择参加什么比赛才能得到更理想的的名次? 即要能根据个体的实际情况对比赛成绩做出较为准确的预测。以上的任务虽然是抽象具体的,但是意义却是非常明确而重大的。不过此类比赛却具有以下的特征和难点: (1) 一个队伍中参加比赛的每一个参赛个体都要有自己相对独立的模型。而针对某一方面的影响因素,或许它对参赛个体 A 有非常明显的影响,而对参赛个体 B 的影响却很小甚至没有任何影响; 很明显在这种情况下对于最后比赛

收稿日期:2011 -01 -15;修回日期:2011 -03 -01.
作者简介:朱文富(1971 -),男,重庆潼南人,讲师,从事体育教育研究.

成绩的预测,不同的个体由于个体自身的差异性导致无法沿用同样的模型来进行最后结果的预测。(2) 每个个体在某一具体比赛中最后取得的成绩也与其临场状况有关,也与其近阶段比赛的历史成绩有一定关系。临场状况在这里指的是个体自身的各方面因素和此次比赛赛场的构成、现场气氛等可以影响比赛成绩的各种客观因素整体所构成的情形。但这不包括由于人为因素故意降低比赛成绩等情况,因为这是无法用科学的方法来进行分析和预测的。(3) 影响成绩的关系错综复杂,而且因素可能很多。(4) 它取值不仅是基于模糊的,而且是主观判断的和定性概念的属性,甚至有些还可能是非常重要属性。例如,身体状况问题、情绪问题等等。(5) 时间的发展对参赛个体成绩的预测也有影响。例如: 某一运动员从高峰期开始走向低峰期或反之,数据都会发生变化。此外,因为每次比赛参赛个体的模型具有多样性多样性,所以就会出现这样一种情况: 一些参赛个体的数据会经过一段时间的积累才会得到相对准确的结果。

2　求解问题的方法和策略

2.1　方　法

2.1.1　可以用关联规则探索或者是确定影响因素

此方法主要包括两种类型: 发现型挖掘和验证型挖掘。其中发现型挖掘一般是用于未被注意到的、发现新的或特定于某个体的影响因素。验证挖掘用于经验所认定的或已由专家验证认定的因素。由于每次参加比赛的群体具有多样性,所以最后得出的数据的可用性就不一定会很充分,而且又因为时间性相对来说很强,时间稍微早些的数据或许就已经不能再用来研究。所以,每当在预测建模的时,为了在多种影响因素中选取最主要的因素,就需要对已经发现的的影响因素进行合理的调整,必要时还要进行重新挖掘。而且,在挖掘的时候,为了选取主要的影响因素,要排除过杂过多的情形,此时置信度则可偏大些,在预测建模和影响因素挖掘之间构成一个优化的循环的过程。

2.1.2　用神经网络方法建立临场状况的描述和分类模型

把某一参赛个体的临场状况所表现出来的情况划分为不同的级别,例如很优、良、中、差、很差等等。

2.1.3　成绩的预测

预测成绩: $s = a_1 s_1 + a_2 s_2 \quad a_1 + a_2 = 1$,其中 $s_1 = E(I)$ 即是参赛个体 I 在同一比赛所能取得的成绩的数学期望值,但必须保证参赛个体是在最近的一个时间段 P 内; s_2 是由回归方程所获得的成绩,该成绩是预测成绩预测。参数 P 视具体的个体情况和实际问题而定,也就是说由于个体自身情况不同或具体出现的问题不同,P 的取值就可能会出现变化。参数 a_1 和 a_2 可根据方程式结果进行调整,其调整的具体方法和重要意义将会在本节稍后进行进一步论述。

由上可以得出一个结论,一个个体在即将要参加的同种比赛中预测取得的比赛成绩会与他近一段时间的所取得的成绩相差不大。例如,一个世界冠军级别的运动员和一个平时成绩非常一般的运动员进行比赛,世界冠军在紧接的比赛中被成绩平平的运动员打败的几率是相当小甚至是没有这种可能的。所以,由于 $E(I)$ 这一参考值是相对稳定的,所以 $E(I)$ 就成为预测比赛结果的一个比较科学而又有用的参考值。而回归模型则是通过挖掘影响因素来建模,是利用通过挖掘已经得到的各种影响因素,来通过逐步的回归,分析相关系数和检验其显著性,从而进一步明确最主要的因素,最后分析出它们之间所存在的相互联系,在必要时还要挖掘相关联规则,然后再重新建立出模型等。采用回归方法,要考虑到对预测值的连续性、精度要求以及计算的效率等问题。

2.1.4　对有关比赛信息的数据进行存储

对每一参赛者的每一比赛,增加存储参数由神经网络模型得到的临场状况级别 L,预测成绩(包括 s_1、s_2、s 和名次),以及 a_1、a_2 值。

2.1.5　模型的评估

以排名为最终标准,以成绩为参考标准。假如能够做到实际取得的成绩与预测的成绩完全相符,实际比赛的排名和赛前预测的排名相符,这当然是最理想的。可是,从另一方面来讲要达到这种很理想的准确

程度是非常不容易的的。当把比赛的成绩来作为预测的直接结果时就具有不受参赛群体数量多少的限制,具有较好的可区分度和灵活性,但是要做到高准确性和高精准度却是非常困难的。此外,对于某些比赛来说,更加看重,更加关心的结果可能会是比赛的名次。

2.1.6 修正或重建模型

在每场比赛结束后,若把预测的成绩设为 S,参赛者 I 的实际比赛成绩为 T,那么当 $|S-T|>\sigma$,即实际值与预测值之间的误差超出某个允许的范围 σ 时,由方程组: $\begin{cases} a_1+a_2=1 \\ s_1a_1+s_2a_2=T \end{cases}$,参数 a_1 和 a_2 须重新确定其取值(最初始,取 $a_1=a_2=\frac{1}{2}$,也可由具体问题而定)。

由该方程组的增广矩阵的秩: $\left[\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & S_1-S_2 & T-S_1 \end{array}\right]$,可分析: 当 $s_2-s_1\neq 0$ 时,方程有解,而且是唯一解。这种情况可按求得的参数 a_1,a_2 值来进行调整; 当 $s_2-s_1=0$ 且 $T-s_1=0$ 时,此方程有无穷个解,这种情况是理想的,即 $s_1=s_2=s=T$,此时的参数不用作调整; 当 $s_2-s_1=0$,但 $T-s_1\neq 0$ 时,方程无解,此时可以进行以下处理: 让 $\varepsilon=\frac{T-S_1}{\text{临场状况级别总数}}\times 2$,这时的状况神经网络 IF 输出值 L 与 s_2 值相符和,也就是 L 值在中等级别以上时且 $s_2>T$ 的情况,或者 L 值在中等级别以下时且 $s_2<T$ 的情况

THEN 说明 s_2 值基本上可以用,s_1 值的误差比较大,此时应该调高 a_1,把 a_2 调低,让 $s'_1=T+\varepsilon\times$(L 值与中等级别之间的级差),最后解方程组: $\begin{cases} a_1+a_2=1 \\ s'_1a_1+s_2a_2=T \end{cases}$ 获取可变参数 a_1 和 a_2 值。

ELSE 说明 s_1 值基本上可以用,s_2 值有比较大的误差,此时应根据实际情况把 a_2 值调高,把 a_1 值调低,调整出 s_2 的值,让 $s_2=T+\varepsilon\times$(L 值与中等级别之间的级差),然后解方程组: $\begin{cases} a_1+a_2=1 \\ s'_1a_1+s_2a_2=T \end{cases}$

获取可变参数 a_1 和 a_2 值。

例 设 $s_1=s_2=86$,$T=100$; 状况级别 NN 的输出 = '差'。因此 $\varepsilon=4$,$s'_2=100+4\times 2=108$,从该方程式求解得出 $a_1=\frac{4}{11}$,$a_2=\frac{7}{11}$。为了让精度更高,a_1 和 a_2 最好用分数表示,以此来避免因小数截断而造成的误差。

参数 a_1 和 a_2 除了对预测进行修正和作为 s_1 和 s_2 的置信比度外,另一方面来说它还具有以下的作用和重要的意义; 如果出现两个值当中的其中一个值始终持续维持在某一低水平,这种情况就表明与个体参赛成绩相关的预测值是不准确的,出现此类情况后就说明该模型就要进行改进,在必要时还需要重新建模。此外,对 a_1 值的变化还具有另外一效用即"趋势发现"的效用,其值能显示出此位参赛个体的发展在近期是提高了已进入高峰期,还是由于某种原因下跌而进入低潮期。很明显,对 a_1、a_2 值的处理是便捷方便捷灵活的,通过系统就可以完全自动实现。虽然,a_1,a_2 值和状况级别 L 值的存储代价不是很高,但是它可以用于模型的进一步改进及重建。

2.2 基本策略与原则

根据以上在实际应用时出现的问题及其特征,在此将利用以下两项原则和三条策略来解决此类问题。原则: 一定要能对模型的重建给出建议和自动的给与提示,并且能为模型重建给出有用的信息; 必须能够方便、灵活且自动地对已有的模型进行修正。策略: 对付主观、模糊的概念属性时用模糊的逻辑。经研究发现,处理复杂问题时采用多方法、多模型是一种有实用性的策略,而且预测精度还可以得到很大程度的提高[3-4]; 而处理主观、含糊的语言变量是运用模糊逻辑则是非常有效而成熟的[5]。以多模型、多方法的组合/结合提高预测的准确性并对付问题的复杂性。当把轻量级模型应用于每个参赛个体时,就必须要用自己独特模型所附带的对资源的特殊要求和效率问题。

2.3 应用实例

把以上方法有效地应用于某一地区的赛车比赛中,首先确定出 8 个对比赛有影响的因素,它们分别为:

决定性因素(赛程) ,赛车本身因素(重量、排量、轮胎性能) ,其他影响因素(赛车手排位、车辆出发排位、场地的性质、天气情况等) 。问题的客观实际就基本上被精准而简练的反映出来。运用建模方法对此种比赛的 50 场比赛结果进行预测,预测的名次与实际名次相差在五名以内的准确率大概为 85% 。具有相当高的准确率。另外有专家也曾经把以上方法应用于赛马娱乐项目,预测所得出的名次与实际比赛的名次相差在三名以内的概率约为 75% 。但是,由于有一些商业方面上的特殊原因(比如其他预测系统的预测: 马评家的观测等) ,此类比赛的结果还不能与其他的同类预测相比较。另一方面,也考虑到一些参赛马匹近期的真实历史资料不太容易获取和马匹的不确定性等问题,这就给预测增添了很大的难度,所以由于各方面的原因,该预测的准确率也还是可以接受的。

3　结　论

通过对几个参数进行简单的分析处理,最后可以方便、自动地修正已存在的模型、完善,把多模型的、轻量的和多种技术的组合/结合作为策略; 研究问题的主要特征而又兼顾到该问题的各个方面,对非对垒式比赛类给出了一个分析以及预测的 KDD 建模方法,对比赛结果作出预测,并对每个参赛者的比赛成绩有影响的各种重要因素进行分析和确定。同时参赛群体许多客观的问题也通过 KDD 建模方法得到了较好的解决,包括多种多样的参赛群体、复杂的影响因素以及预测上的困难等一系列问题。此外,KDD 建模方法不仅能够主动对模型的改进或重建给出合理的建议,还可以为模型的改进或重建提供有用的信息。

参考文献:

[1] 陆伟,吴朝晖. 知识发现方法的比较研究 [J] . 计算机科学,2000,27(3) : 80-84

[2] 欧阳为民,郑诚,张燕. 国际知识发现与数据挖掘工具评述 [J] . 计算机科学,2001,28(3) : 101-108

[3] 刁力力,胡可云,陆玉昌,等. 数据挖掘与组合学习,计算机学习 [J] . 2001,28(7) : 73-78

[4] 韩宏,杨静宇. 多分类器组合及其应用 [J] . 计算机科学,2000,27(1) : 58-61

[5] SMITH E, ELOFF J. Cognitive fuzzy modeling for enhanced risk assess-ment in a health care institution [J] . IEEE Intelligent Systems,2000,15(2) : 69-75

[6] NARENDA K S, MUKHOPADHYAY S. Adaptive Control ofnonlinearmulti - variable systems using neural networks [J] . NeuralNetworks,1994,7(5) : 737-752

[7] 张德培,罗蕴. 应用概率统计 [M] . 北京: 高等教育出版社,2000

[8] 焦李成. 神经网络系统理论 [M] . 西安: 西安电子科技大学出版社,1990

Research on Predication Model for Sports Match Results

ZHU Wen-fu

(School of Physical Education, Chongqing Technology and Business University, Chongqing 400067, China)

Abstract: Through computer data decision analysis method, according to sports match based on the ranking of individual achievement, modeling methods are set up based on analyzing competition achievement and predicating competition results to analyze and determine the important factors influencing competition results of each athlete and the model is tried to forecast the match results.

Key words: KDD; match; predication; modeling

责任编辑: 代小红

校　　对: 罗泽举