**基于最大熵的团体竞技比赛胜负的预测**

数学与计算科学院，湘潭大学，湘潭，中国

**摘要：**

利用最大熵原理，将影响某种团体竞赛的不同的因素对比赛结果影响的权重，计算出来，构造出符合该团体赛事的最佳条件概率模型。通过对团体赛事的最近几场比赛中因素的具体状况，计算出符合该赛事的最佳平稳周期，以每个周倜为单位，将该周期中影响因素的参数带入模型中，计算出团队赛比赛的胜率。

**关键字：**

最大熵，团体比赛，平稳周期，期望，胜负，预测

**引言：**

随着互联网的发展和人们对于各种体育竞技赛事的关注度的提高，像欧洲足球五大联赛、美国职业橄榄球联赛、美国职业篮球联赛等团体体育的竞技赛事已经不只受到当地居民的关注，这些赛事已经普及全球。各大型博彩公司、大型网站纷纷对赛事的结果进行分析与预测，比赛关注者根据自己的主观偏好和网站给出的分析对一场比赛的输赢预测。团体赛事的的预测也早有好多学者给出过不同的方法，但预测的效果不是十分理想。最大熵原理指出，当我们需要对一个随机事件的概率分布进行预测时，我们的预测应当满足全部已知的条件，而对未知的情况不要做任何主观假设。本文将最大熵原理运用到团体的赛事中，通过实验数据的结果分析能得到一个有更高准确率的结果。

**正文：**

**一 最大熵模型的介绍**

1948年C.E.Shannon首次提出了信息熵的概念，把熵作为一个随机事件的“不确定性”或信息量的度量。



式中为信息熵，为随机事件出现的概率，为常数。

1957年E.T.Jayne提出了一个准则：当根据部分信息进行推理时，必须选择这样一组概率分配，它应具有最大的熵，并服从一切已知的信息，这一准则被称为最大熵原理。最大熵原理指出，当我们需要对一个随机事件的概率分布进行预测时，我们的预测应当满足全部已知的条件，而对未知的情况不要做任何主观假设。在这种情况下，概率分布最均匀，预测的风险最小，因为这时概率分布的信息熵最大，所以人们称这种模型叫“最大熵模型”。“最大熵模型”属于指数模型的一种：它将每个信息源视为一组约束条件，在满足所有约束条件的一组概率分布中，寻找其中熵最大的分布。它的特点是能以满足限定条件下的熵值最大化为准则，对各种不用类型的特征训练出一组对应的特征权值，然后通过线性组合，把它们整合到一个统一的模型中。其基本原理的形式化描述如下：

条件信息熵，即模型中的目标函数，定义如下：



其中为类别比标记，为标记时出现的相关信息。



其中表示特征在模型中的期望值，表示特征在训练集上的经验期望值，二者相等就表示目标模型要尽可能吻合于实际训练数据，其中，和的定义如下：







所求解就是式（1）在满足式（2）的条件下，条件概率的最大值。

对（1）（2）使用拉格朗日变换求出满足条件极值得概率如下：



其中是归一化因子：

最大熵模型采用指数的形式来计算条件概率。由公式(5)可看出，最大熵算法的学习过程就是调整的过程。称为特征，的权重。可以通过GIS(Generalized Iterative Scaling)算法分配[1]或者IIS()算法分配[2]。

**二 将最大熵原理运用到团体赛事中构造最佳的模型**

**2.1** NBA是篮球的最高的层次，是受关注度和普及度最高的项目之一。其最大的特点就是不可预测性，因为高强度，快节奏的比赛，往往在一分钟甚至一秒钟的时间内，比赛的结果都有可能不同，比如麦迪的34秒13分的神奇纪录。影响胜负的因素很多，虽然说实力因素是主导因素，但是临场的发挥，心理因素，运气成分，一些突发的情况等也是决定比赛胜负的重要因素，而这些因素也是随机，不确定的，导致了竞技比赛的赛果难以预测。

近年来，体育竞技的分析已经演变成为一个重要的领域。国内外有很多相关研究文献,尽管这些研究文献较多, 但绝大多数是使用文献资料、比较分析和录像观察等方法进行的研究，参考文献[2][3][4], 部分也使用了一些较简单的描述性统计分析方法，比如：Logistic模型找出主要影响因[4]、Fisher 判别分析[5], 主成分分析/聚类分析/相关分析[6],Logistic 和Bayes 模型[7]仅有少数使用统计建模方法进行实证分析。由于分析的手段不同、角度不同, 尚有许多问题没有涉足, 尤其是基于统计建模方法对于NBA比赛胜负的影响因素进行较全面的分析, 有待进一步深入。为此, 本研究综利用最大熵原理, 通过大量的训练数据的反复试验，试图较全面地考察NBA比赛的影响因素，并分析不同因素对比赛结构影响的大小。

**2.2**样本表示与特征提取

对1995～2005年20个赛季的NBA全部比赛的各项技术统计进行了归纳。所涉及的篮球技术指标包括：出手投篮数、投篮命中率、三分球投篮数、三分球命中率、罚球次数、罚球命中率、前场篮板(进攻篮板)、后场篮板(防守篮板)、助攻、盖帽、抢断、犯规、失误、球队价值(values)、工资总额(pay)、教练执教能力(coach)和核心运动员能力(player)，两球队是之间的距离（meters）十八项，在研究因子上力求更加宽泛，使得研究结果更具说服力和实效性。将十七项指标分为球队内部因素和非球队内部因素两大方面的影响因素。其中球队内的影响因素包括：出手投篮数、投篮命中率、三分球投篮数、三分球命中率、罚球次数、罚球命中率、前场篮板(进攻篮板)、后场篮板(防守篮板)、助攻、盖帽、抢断、犯规、失误、核心运动员能力(player)。球队外的影响因素：球队价值(values)、工资总额(pay)、教练执教能力(coach)，两球队是之间的距离（meters）。

鉴于主客场各项指标的差异, 以主场比赛的胜负为因变量Y, 20个赛季的技术指标为自变量X, 自变量还加入主客场因素(field)作为虚拟变量(主场值为2, 客场值为1), 虚拟变量与前面技术指标[出手投篮数、投篮命中率、三分球投篮数、三分球命中率、罚球次数、罚球命中率、前场篮板(进攻篮板)、后场篮板(防守篮板)、助攻、盖帽、抢断、犯规、失误、球队价值(values)、工资总额(pay)、教练执教能力(coach)和核心运动员能力(player)，两球队是之间的距离（meters）]进行乘积的变量也作为自变量进入模型。

**2.3**训练得最佳模型

将训练集（X,Y）的数据代入条件信息熵模型中



其中为技术指标，为技术指标时的主队的获胜的事件。





所求解就是式（1）在满足式（2）的条件下，条件概率的最大值。

对（1）（2）使用拉格朗日变换求出满足条件极值得概率如下：



通过GIS(Generalized Iterative Scaling)算法计算出的值，如下：

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

从而得出最佳模型

**三 平稳周期的选取**

对于影响比赛结果的指标不变的情况下。系总是一个周期的循环。最为预测我们需要比赛前的某几场比赛的各项指标的参数作为即将比赛参数的一个预测值。

现在我们寻找最能体现即将比赛的最可能的各指标的数据。

首先，我们将最近n(n=2,3,4,5,6,7,8)场的比赛的十八项指标的数据分别求其期望（i=1,2,...,18），再将这十八项指标进行方差求和。如下表：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

在此，我们定义：最小的情况下，n的取值就为NBA比赛的赛事的稳定周期。

**四 数据带入分析**

取赛前最近的最佳稳定周期中的各项指标的均值代入最佳模型中，求出本场比赛的主队获胜的概率，我们将训练数据集作为测试数据集进行测试检验。可以得出该模型的预测准确率为： 。可见模型预测效果较好。

**结论：**

通过将数据带入模型中进行检验，分析可知，将最大熵原理的方法应用到团体竞技的体育赛事中是可行的。实例应用表明，该模型算法简便，能准确标定参数，且有相对较高的预测准确率。

**参考文献：**

[1]Ratnaparkhi A．Maximum entropy models for natural language am—biguity resolution[D]．University of Pennsylvania．1 998．

[2]NBA2004～2005赛季常规技术统计与比赛胜负关系的研究

[3] ２０１０－２０１１赛季ＮＢＡ总决赛进攻技术研究

[4]２０１１－２０１２赛季ＮＢＡ总决赛技术统计分析

[4]影响NBA 常规赛胜负的Logistic分析

[5]对NBA 常规赛比赛胜负影响因素及Fisher 判别分析

[6]NBA球队战绩影响因素的统计分析

[7]NBA 季后赛成绩分析及预测：Logistic 和Bayes 模型