基于决策树 C4.5 算法的足球赛事预测

闵 芳，杨功廷，张 昱

（南京航空航天大学 金城学院，南京 211156）

摘要：随着人们对各类足球赛事关注度的不断提高，以及赛事分析预测和足球博彩行业的不断发展，较为精确的足球比赛的胜负规律分析具有一定的商业价值和推广价值。首先对足球比赛的历史数据进行采集，通过各类足球比赛建立其相应的数学模型；进而将采集到的比赛历史数据应用在该数学模型之上，这将可能使得影响比赛走势的因素数字化。并且找到一个从相关比赛数据到比赛胜负平的映射关系，最终将这个映射关系应用于未知比赛的分析上，分析结果使用 C4.5 算法实现。

关键词：C4.5；数据挖掘；足球赛事胜负分析

中图分类号：TP391 **文献标志码**：A **文章编号**：1671－1807(2014)06－0094－03

在数据挖掘迅猛发展的今天，越来越多的事物得以被更加科学的分析，这同时也揭示了许多隐藏在事物背后的模式和规律，例如经典的啤酒与尿布的案例。而本文重点研究的是足球比赛中球队的胜负与之前历史战绩的规律。本文主要研究的是记录比赛信息的数据库设计和基于 C4.5 算法的胜负分析算法的设计，最终以网页应用的形式来表现[1]。

1 整体设计

对于足球赛事胜负分析系统来说，该系统应该被设计成一个可以独立运行，自动获取更新比赛的各类信息，并由此更新分析算法，生成分析结果的系统。

2 足球赛事预测模型

2.1 球队进球率模型

进球率是客观反映球队进攻实力的数据，即一只球队每场比赛的平均进球数。其计算公式为：

进球率＝总进球数/总比赛场数 (1)

当获得交战双方的进球率后，即可将两队进球率相减（一般情况为主队进球率－客队进球率），可以得到进球率之差，而两队的进球率之差在很大程度上可以影响到球队的胜负走势。

2.2 球队积分模型

球队积分是客观反映了球队的成绩，而近几轮的球队积分则是球队近期的状态很好的体现。首先必须说明对于一支球队来说，赢得一场比赛，可以获得三分；平一场比赛，可以获得一分；输掉一场比赛，则一分不得。对于球队积分模型来说，通过计算每支球队的积分之后，反应球队状态。近五轮积分越高，则球队状态越好，在下一轮的比赛中，取得胜利的可能性越高。为了使该数值能更加准确反应球队的竞技状态，在该模型的基础上，再加上各支球队的胜率的权重，还需要特别指出的是这里的胜率是指主场球队的主场胜率和客场球队的客场胜率，这样可以使该模型更加严谨。本系统通过比较近五轮积分和球队胜率的信息增益率后，发现其比例约为 6∶4，而又因为近五轮积分的数值分布为 0 至 15，因此，这里将近五

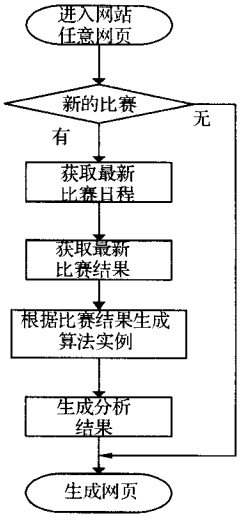

图 1 系统流程图

收稿日期：2014－03－22

作者简介：闵芳（1980—），女，江苏宜兴人，南京航空航天大学金城学院，讲师，硕士研究生，研究方向：数据处理。

轮积分加上两队的主场或客场胜率乘以系数 10，得到一个数值分布为 0 至 25 的指数，在此先称之为胜利指数。同理，对即将进行比赛的两支球队来说，求其胜利指数之差。从纯数据的角度上来说，若该差值大于零，则主队赢得比赛；若该差值等于零，则两队打成平局；若该差值小于零，则客队赢得比赛。

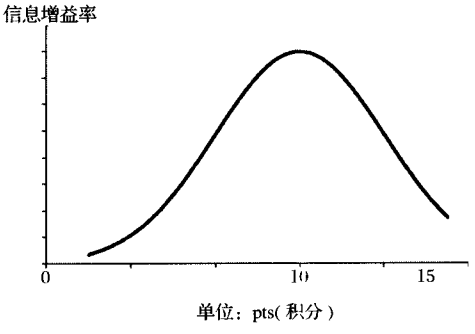

图 2　近五轮积分的信息增益率曲线

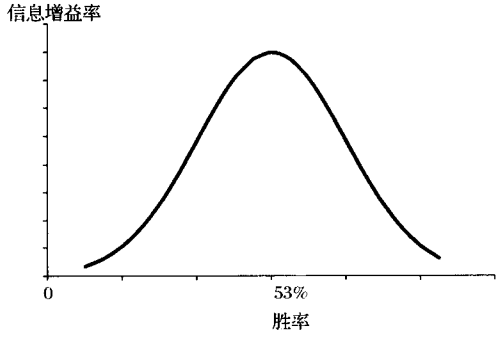

图 3　球队胜率的信息增益率曲线

球队积分公式为

胜利指数＝近五轮积分＋球队胜率＊10　　(2)

从以上关于近五轮积分和球队胜率对于信息增益率的正态分布图中。不难看出，对于近五轮积分这一数值来说，当近五轮积分达到 10 的时候，信息增益率达到峰值。可以理解成 10 是区分两个不同分类的一项阀值。而对于胜率来说，当胜率达到 53%时，信息增益率达到峰值，也就是说 53%也是区分两个不同分类的一项阀值。最后再根据阀值的分布情况，具体判断取值范围[2-3]。

3　数据挖掘 C4.5 算法及其应用

3.1　C4.5 算法简介

C4.5 算法的实质就是由样本集生成决策树的过程。结果，该算法生成了一个决策树形式的分类器；决策树是具有两类节点的结构：叶节点表示一个类；决策节点指定要在单个属性值上进行的检验，对检验的每个可能输出都有一个分支和子树[4]。

3.2　C4.5 预测模型

基于上述介绍的两个数据模型，可以一个这样的数据集，该数据集含有两个属性值，分别为进球率和近五轮积分，通过这两个属性，可以划分成胜负平三类。通过 C4.5 算法，以这个数据集为基础生成一个决策树。不难想象出，该数据集的分类分布可以看成在一个二维空间中的两个区域。在此需要特别强调，因为在足球比赛中，平局是个难以预知的中间状态，因此，在应用 C4.5 算法时，不把平局当作成一种分类，即所有分类为平局的数据，暂时不对其作出处理[5]。

通过 C4.5 算法生成的决策树，去除掉极端情况，可以得到如下分布图。

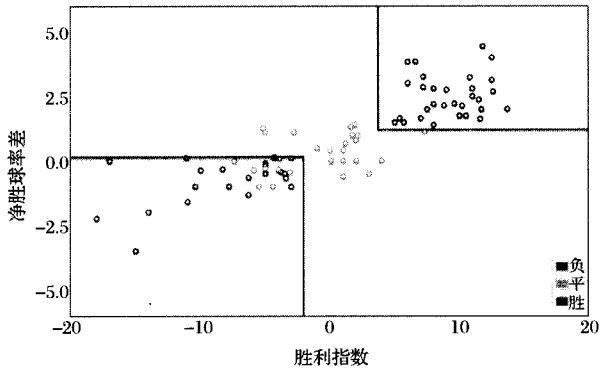

图 4　胜负分布情况图

3.3　对胜负平的分析预测

对数据集使用 C4.5 算法生成决策树后，系统可以获得一个大致的胜负分析标准，若具化在二维坐标系中，可以理解成一个范围。比如在比赛前，通过通过计算得到的进球率差值及胜利指数差值在胜利的范围内，可以理解成这场比赛主队有极大的可能取胜，而计算出这个可能性有多大，便是本算法的最后一步。

对于待分析预测的比赛，先计算出比赛的进球率和胜利指数，根据数值分布分别处理：

在描述处理步骤之前，再次声明一下所需要用到的参数。

所有分析属于失败确胜利的数据占所有分析属于失败数据的比例 L；所有分析属于胜利却失败的数据站所有胜利数据的比例 W；划分胜利边界的交点 $A(X_1,Y_1)$；划分失败边界的交点 $B(X_2,Y_2)$；点 $C(X_3,Y_3)$为点 A 与点 B 的中点；该数据为点 $D(X_4,Y_4)$设胜利的权重为 W_1，平局的权重为 W_2，失败的权重为 W_3；$S=W_1+W_2+W_3$。

1. 对于分布在胜利的数据的处理：

$$W_1=\frac{W*100+X_1+Y_1-X_4-Y_4}{S},$$

$$W_2=\frac{X_3+Y_3-X_4-Y_4}{S},$$

$$W_3=\frac{(1-W)*100+X_4+Y_4}{S};\qquad(3)$$

2. 对于分布在失败的数据的处理：

$$W_1=\frac{L*100+X_4+Y_4-X_2-Y_2}{S},$$

$$W_2=\frac{X_4+Y_4-X_2-Y_2}{S},$$

$$W_3=\frac{(1-L)*100+X_4+Y_4-X_2-Y_2}{S};\qquad(4)$$

3. 对于分布在胜利或者失败范围之外的数据的处理：

$$W_1=\frac{|X_4-X_1|+|Y_4-Y_1|+|X_4-X_2|+|Y_4-Y_2|+|X_4-X_3|+|Y_4-Y_3|}{|X_4-X_1|+|Y_4-Y_1|}\Big/S,$$

$$W_2=\frac{|X_4-X_1|+|Y_4-Y_1|+|X_4-X_2|+|Y_4-Y_2|+|X_4-X_3|+|Y_4-Y_3|}{|X_4-X_3|+|Y_4-Y_3|}\Big/S,$$

$$W_3=\frac{|X_4-X_1|+|Y_4-Y_1|+|X_4-X_2|+|Y_4-Y_2|+|X_4-X_3|+|Y_4-Y_3|}{|X_4-X_2|+|Y_4-Y_2|}\Big/S;\qquad(5)$$

该算法对胜负平的分析，主要是基于 C4.5 生成决策树时所产生的阀值的数学优化。该概率数值根据与阀值点的相对位置大小来决定概率的大小。如情况 1 中的胜利概率，若一场比赛的分析数据处于在分析的二位空间中处于胜利阀值点的左下端，则越偏离阀值点，该场比赛的胜利几率越大，相对平局和失败概率也越小[6]。

4 实验数据

比赛结果经过一定处理，在用经过改进适用于足球比赛的 C4.5 算法分析，以图 5 所示圈状分布图和图 4 的点阵图的形式给出。

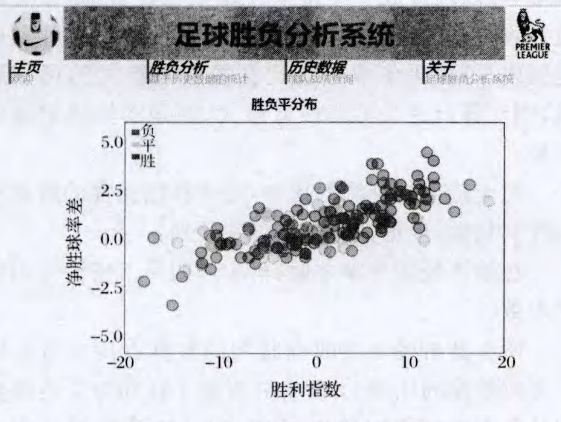

图 5　比赛结果圈状分布图

在实现系统应用的基础上，系统基于历史比赛数据对足球比赛的分析的结果，也与真实结果较为相似。

表 1　足球比赛分析准确度测试表

	测试数据	预期结果	实际结果	与预期结果是否一致
第 25 轮比赛数据	利物浦 vs 阿森纳	胜利	胜利	完全一致
	维拉 vs 西汉姆	胜利	失败	不一致
	切尔西 vs 纽卡	胜利	胜利	完全一致
	水晶宫 vs 西布朗	胜利	胜利	完全一致
	桑德兰 vs 赫尔城	胜利	平	基本一致
	安普敦 vs 斯托克	胜利	平	基本一致
	诺维奇 vs 曼城	胜利	平	基本一致
	斯旺西 vs 加的夫	胜利	胜利	完全一致
	热刺 v 埃佛顿	胜利	胜利	完全一致
	曼联 vs 富勒姆	胜利	平	基本一致

尽管足球比赛的胜负在一定程度上有规律可循，但也时常出现一些超出规律之外的时间，因此对足球比赛胜负的分析是绝不可能做到 100% 的准确。此外，由于足球比赛中存在平局这一局面，且考虑到该局面的不可分析性，当实际结果为平局时，不能完全说明分析结果与实际结果不一致，而应该是基本一致[7]。

5　结束语

本文针对当前数据挖掘算法大行其道的趋势，结合足球这一项世界热门的运动，对基于数据挖掘的足球赛事胜负分析算法进行了研究与改进，探究了足球比赛历史信息对于足球比赛的规律模式。经过对数据挖掘 C4.5 算法的学习研究和对足球比赛模型的钻研想象，较为深入地掌握了分析足球比赛的一般方法，也对数据挖掘的在应用上有了更深层次的理解。

（下转第 128 页）

的建设实施不但需要图书馆对自身业务及对基于网络的三维虚拟图书馆的概念进行深入理解，而且还需要社会的高度重视及各部门的积极配合。

参考文献

[1] 胡伟斓，潘志庚，刘喜作，方贤勇，石教英．虚拟世界自然文化遗产保护关键技术概述[J]．系统仿真学报，2003(3)：315－318.

[2] RONALD T AZUMA，et al. Recent Advances in Augmented Reality. IEEE Computer Graphics and Applications，2001(21)：1－15.

[3] SONG MEEHAE，MULLER-WITTIG WOLFGANG，CHAN TONY K Y. Reconstructing Peranakan Identities through Digital Heritage[C]//Proceedings of VSMM 2002：124－131.

[4] 赵一鸣，吴署旻，潘志庚．网上3D虚拟商城的设计与实现[J]．系统仿真学报，2003(7)：980－986.

Research on Key Technology of 3D Virtual Library Based on Network

HAN Yan-ping

(Party School of C. P. C Huaihua Committee，Huaihua Hunan 418008，China)

Abstract：In this paper，the concept and construction of Web-Based 3D Virtual Library are studied mostly，and based on IBR (image-based rendering) technique with computer graphics and VR (Vertual Reality) and AR (Augmented Reality)，Web-Based 3D Virtual Library with a realistic environment has been designed. Concerning intellectual property rights and culture heritage protection inside and outside the nation nowadays，it proves that Web-Based 3D Virtual Library becomes a development trend，and that has a promising future.

Key words：network；the 3D virtual；library

(上接第96页)

参考文献

[1] JIAWEI HAN，MINCHERLINE KAMBER. 数据挖掘概念与技术[M]．范明，孟晓峰，译．北京：机械工业出版社，2006：15－16.

[2] 艾芳菊．基于实例推理系统中的权重分析[J]．计算机应用，2005(5)：1022－26.

[3] KENNETH D LAWRENCE，STEPHAN KUDYBA，RONALD K KLIMBERG. Data minin methods and applications [M]. Boca Raton，FL：Auerbach Publications，2008：56－57.

[4] JUAN WANG，QIREN YANG，DASEN REN. An Intrusion Detection Algorithm Based on Decision Tree Technology[J]. Information Processing，2009(18－19)：333－335.

[5] 王宏威．基于决策树的分类算法研究[J]．软件导刊，2007(9)：134－135.

[6] 陈竞艺．基于数据挖掘的入侵检测系统在校园网中的应用[D]．石家庄：河北科技大学，2011.

[7] 谢邦昌．数据挖掘基础与应用[M]．北京：机械工业出版社，2011：78—90.

Forecast Football Match Baseon C4. 5 Decision Tree

MIN Fang，YANG Gong-ting，ZHANG Yu

(Nanhang Jincheng College，Nanjing 211156，China)

Abstract：As different categories of football matches draw the same rapidly growing attention of the public，as the trade of football lottery and the profession of the analysis and prediction of the matches are uprising，there indicates that there must be commercial and promotion value in a relatively precise analysis of the won lost percentage in matches. This thesis researches mainly on the design of the database that keeps an account of detailed football matches' information and on the design of algorithm that analyze the won lost percentage which is based on C4. 5 algorithm.

Key words：C4. 5；Data；mining；football matches forecast system