数据挖掘在足球运动中的应用

成孟金，曹志宇

（沈阳化工学院 计算机科学与技术学院，辽宁 沈阳 110142）

摘 要：数据挖掘是指从数据库的大量数据中揭示隐含的、先前未知的、潜在有用信息的非平凡的过程。使用可视化数据挖掘的技术从足球比赛的数据集中找到模式。这些模式可以在足球比赛中直接或间接地提供有益的见解，并在比赛中运用决策支持系统。

关键词：数据挖掘；可视化；模式

中图分类号：TP274 **文献标识码**：A

Applying data mining techniques to football

CHENG Meng Jin, CAO Zhi Yu

(Computer College, Shenyang Institute of Chemical Technology, Shenyang 110142, China)

Abstract: Data mining is the process of finding new, potentially useful and non trivial knowledge from data. In this paper, visualization techniques has been used to find patterns in datasets from football games. Data mining is the process of finding new, potentially useful and non trivial knowledge from data. Football is the world's first sports and it is a very popular game, it has a rich source of data. In this paper, we will use visualization techniques to find patterns in datasets from football games. If these patterns can provide valuable insight to the people involved directly or indirectly in the match. Application would be the development of a decision support system to be used during the match.

Key words: Data mining; Visualization; patterns

数据挖掘DM(Data Mining)技术在足球运动中的运用潜力是非常巨大的。足球运动起源于英国，它的巨大影响力与日俱增，在世界上已经有超过240万人从事这项体育运动[1]，有着非常丰富的数据资源。

跨行业数据挖掘过程标准CRISP-DM (Cross-Industry Standard Process for Data Mining)是由欧洲几家在数据挖掘应用上有经验的公司共同筹划组织的一个特别小组所提出的，它分为6个阶段，在本文中主要包括3个部分[2]：第1部分，定义商业问题(business understanding)，本阶段的主要工作是针对该课题的目标和需求进行了解确认，针对不同的需求做深入了解，将其转换成数据挖掘的问题，并拟定初步构想去实现该目标。第2部分，数据理解(data understanding)和数据预处理(data preparation)，数据理解阶段以收集数据开始，了解数据的含义与特性，并过滤出所有可能有用的数据，然后进行数据整理并评估数据的质量，把各种不同来源的数据加以整理和归并，以适合数据挖掘技术的使用。第3部分，包括CRISP-DM的建立模型(modeling)阶段，使用可视化的技术来挖掘数据。

1 定义商业问题

通过网站zerozerofootball获得了许多欧洲冠军联赛和许多国家的足球联赛数据，其中主要的2个数据集：(1)在2008、2009年的葡超冠军联赛中，因为它是具有最高的详细程度和水平最低的遗漏值和错误数据。(2)在过去的50年，6个欧洲国家的所有比赛，也包括葡萄牙联赛。

通过所选择的数据集，用数据挖掘技术做探索性工作从而找出它的模式，即可以在足球比赛中直接或间接地提供有益的见解。达到在比赛中运用决策支持系统、对比赛的结果进行预测的目的[3]。

2 数据理解和数据预处理

建立数据库和分析数据，包括一些欧洲国家足球联赛的冠军和比赛的信息，如葡萄牙自从1934年以来的15 382场比赛，英格兰从1888年起的43 730场比赛，西

班牙从 1930 年起的 19 846 场比赛，意大利从 1946 年起的 17 680 场比赛，法国从 1933 年起的 22 702 场比赛，以及德国从 1933 年起的 13 406 场比赛。在这些数据中找出影响最大的数据，像队伍的名字、每场比赛的进球数、失球数和胜利者、所处于的国家和年份、每个联赛中每个队伍的总进球和失球数以及每个队伍所获得的分数与胜、负、平的场次数[4]。

还选择了具有最高的详细程度和水平最低的遗漏值和错误数的联赛，2004、2005 年的葡超冠军联赛，这一年的联赛共包括 18 支队伍、总计 306 场比赛，一共有 711 个入球、裁判出示了 1 771 张牌，这一年的比赛信息还包括每场比赛中的队员、替补、以及比赛的时间和地点，例如知道了联赛中每个球队，就知道了它的总进球和失球数以及每个队伍所获得的分数，同时如果知道了 1 个足球运动员的名字，也就知道了该队员的进球数、获得的牌数、助攻数等。图 1 中所示 FC Porto、Benfica、Sporting 在近几十年的联赛里最后所处的联赛排名。

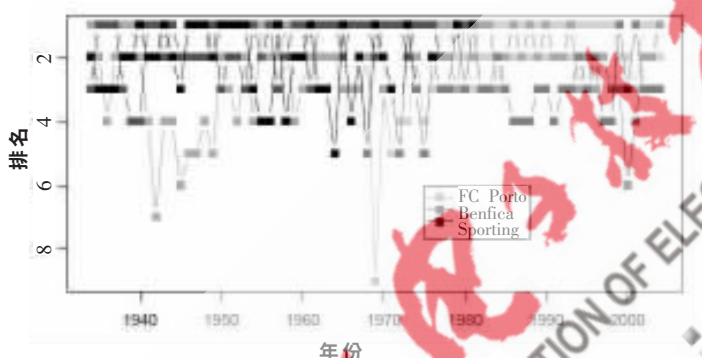

图 1 联赛排名

3 建立模型

数据挖掘的可视化技术是指运用计算机图形学和图像处理技术，将数据换为图形或图像显示出来，并进行交互处理的理论、方法和技术。主要是在相同或相似的数据中给人们一些观察和见解。根据图 1 所示葡萄牙联赛争夺冠军的主要 3 支队伍，通过图 2 可以得到葡萄牙联赛这 3 支队伍获得冠军的分数，并了解这些队伍的变化，也能看出自从 20 世纪 90 年代初改变了规则，即赢 1 场球从 3 分变为 2 分后，FC Porto、Benfica 获胜的次数明显增多了，并且与 Sporting 之间的差距越拉越大。

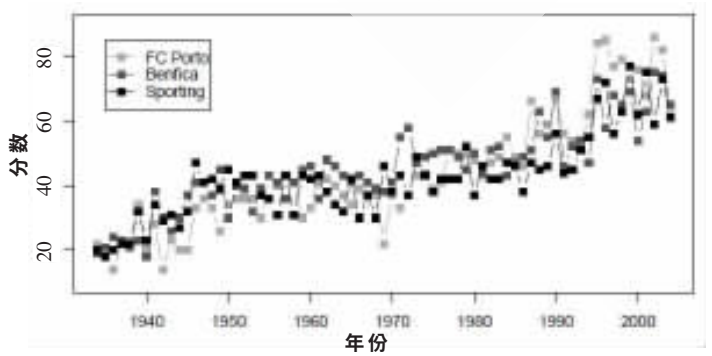

图 2 联赛分数

通过对每一个国家的每场比赛结果加以分析，比赛结果用 2D 的图来表示，不同的黑色阴影表示过去的每年联赛平均每场的得、失球，图 3、图 4 所示为西班牙、英格兰的联赛比较。

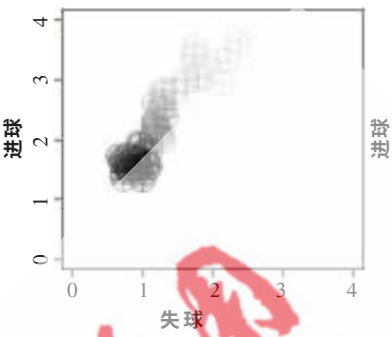

图 3 1930–2008 年西甲联赛比赛结果

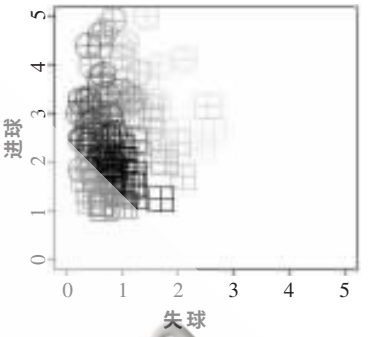

图 4 1988–2008 年英超联赛比赛结果

从对比中可以看出，尽管近几年 2 个国家的比赛结果很相似，但是从总体上和历史上看，英格兰的足球比赛结果有着比较少的变化，而西班牙过去的比赛结果和近几年的结果有着很大的不同。同样，还可以通过数据去衡量 1 支队伍的主客场成绩变化和 2 支队伍更可能出现的结果。例如图 5 所示 Benfica 队的历史主客场成绩(圆表示主场成绩，方块表示客场成绩)，可以看出，近些年该队伍的主场成绩有很大改观。

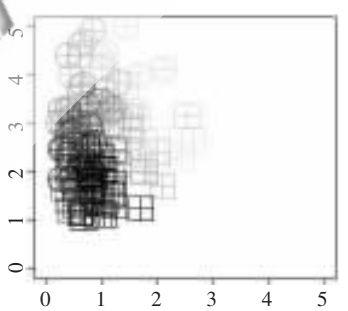

图 5 1934–2008 年 Benfica 的主客场成绩

图 6 所示 FC Porto 对 Benfica 的主场交战记录，每个坐标是比分，从比分的模式可以看出，FC Porto 对 Benfica 的成绩占据优势，平局其次，输球的结果比较少。

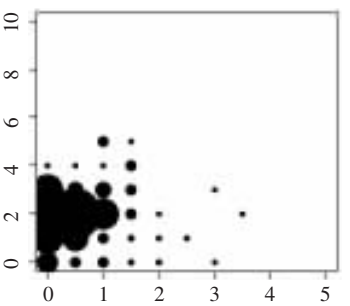

图 6 1934–2008 年 FC Porto 对 Benfica 交战记录

数据挖掘技术是伴随着行业数据量的迅速膨胀和对知识发现的迫切需要所出现的产物，可以实现对足球比赛数据的挖掘，可以更容易得到有根据的模型。但是此项技术作为有效的信息处理和强大的数据分析工具还需要体育专业人员和有经验的分析人员共同完成[5]，该领域有着非常广阔的发展前景。

参考文献

[1] BHANDARI I, Advanced scout: Data mining and knowledge discovery in NBA data[J], 1997.

[2] 郝丽，刘乐平，王星.数据挖掘在体育统计中的应用[J].东华理工学院学报，2004，23(2)：92-95.

[3] 韩凤芝，杜修平.数据挖掘在职教体育教学中的应用[J].中国职业技术教育，2004(31)：38-39.

[4] 隆益民.数据仓库与数据挖掘[J].现代电子技术，2000(10)：70-73.

[5] 杨双燕，赵水宁.体育数据分析中数据挖掘技术的应用[J].浙江体育科学，2003，25(4)：49-51.

(收稿日期：2009-07-04)