doi:10.3969/j.issn.1673-9833.2012.05.016

基于C4.5算法的健身俱乐部会员数据挖掘研究

邓 程，朱艳辉，杜 锐，鲁 林

（湖南工业大学 计算机与通信学院，湖南 株洲 412007）

摘 要：以已投入使用的健身俱乐部管理系统为背景，提出了用C4.5决策树分类算法对健身记录进行数据挖掘。通过该方法找出俱乐部在有效期内的会员的年龄段、性别、会员卡类型和参与健身时间段的规律，提取特定时间段内参与健身的会员特征。实验结果表明：将该分类规则应用到会员管理系统中，可以辅助健身俱乐部的管理者有针对性地制定营销方案。

关键词：C4.5算法；健身俱乐部；会员分类

中图分类号：TP301.6 **文献标志码**：A **文章编号**：1673-9833(2012)05-0071-05

Study on the Fitness Club Membership Data Mining Based on C4.5 Algorithm

Deng Cheng，Zhu Yanhui，Du Rui，Lu Lin

(School of Computer and Communication，Hunan University of Technology，Zhuzhou Hunan 412007，China)

Abstract：Based on the developed and used management system of fitness club, puts forward the application of C4.5 decision tree classification algorithm for the fitting-record data mining. By this method finds out the members' age, gender, membership card types and the participating time rules in their club current validity period, and extracts fitness participate member features at a specific period of time. The experimental results show that the classification rules applying to the member management system can help health club managers make targeted marketing scheme.

Keywords：C4.5 algorithm；fitness club；membership classification

0 引言

随着各种计算机管理软件的普及，其数据库中积累了大量的业务数据。数据挖掘的目的就是找出蕴含在这些数据中的组织业务活动规则，以辅助管理者做出正确决策。将数据挖掘技术应用在健身行业的数据中，能更加客观真实地反映出健身俱乐部当前会员的动态特性，找出不同会员群与健身时间之间的关联规则，进而能更科学地安排健身课程时间，有针对性地制定营销方案。

决策树是一个类似于流程图的树结构，其中每个内部结点表示一个属性测试，每个分支代表一个测试输出，而每个叶子结点代表类或类分布，即决策树根据不同的特征，以树型结构表示分类或决策

收稿日期：2012-08-30

基金项目：国家自然科学基金资助项目（61170102），湖南省自然科学基金资助项目（10JJ3002），国家社科基金资助项目（12BYY045），教育部人文社会科学研究青年基金资助项目（09YJCZH019），中国包装总公司科研基金资助项目（2008-XK13）

作者简介：邓 程（1977-），男，湖南株洲人，湖南工业大学硕士生，主要研究方向为软件工程和数据挖掘，E-mail：29794507@qq.com

通信作者：朱艳辉（1968-），女，湖南湘潭人，湖南工业大学教授，主要从事智能信息处理和信息检索方面的研究，E-mail：swayhzhu@163.com

集合，产生规则和发现规律。决策树具有易于提取显式规则，计算量相对较小，能显式重要的决策属性，分类准确率较高等优点，因而得到广泛应用。当前比较主流的决策树算法有CLS（concept learning system）、ID3、CHAID（chi-squared automatic interaction detection）、C4.5、CART（classification and regression tree）、SLIQ（supervised learning in quest）、SPRINT（scalable parallelizable inndution of decision tree）等。

在株洲某健身俱乐部，会员健身管理系统已使用近3年，积累了大量的客户数据资源。根据本文原始数据的特性和挖掘任务的性质，提出了将C4.5算法对该会员数据进行数据挖掘，找出不同会员群与健身时间之间的关联规则。

1 C4.5算法介绍

C4.5算法是对ID3算法的改进，其优势是将信息增益比作为标准来选择分支属性，这样不仅可以处理离散属性，也能处理连续属性。

1.1 选择属性

C4.5算法选择具有最高信息增益率的属性作为测试属性。信息增益率的计算公式为

$$G_{\text{Gain\_Ratio}}(S,A)=\frac{G_{\text{Gain}}(S,A)}{S_{\text{SplitInfo}}(S,A)}, \qquad (1)$$

式中：分裂信息$S_{\text{SplitInfo}}(S, A)$表示按照属性$A$分裂样本集$S$的广度和均匀性，即

$$S_{\text{SplitInfo}}(S,A)=-\sum_{i=1}^{c}\frac{|Si|}{|S|}\log_2\frac{|Si|}{|S|},$$

其中$S1,\cdots, Sc$是属性A的c个不同值分割样本集S而形成的c个子集；

$G_{\text{Gain}}(S, A)$为属性A的信息增益。

1.2 处理连续属性

选择某节点上的分枝属性时，按照该属性本身的取值个数进行计算。对于某个连续属性Ac，假设在某个结点上的数据集的样本数量为t，将连续属性Ac离散化处理[1]，具体实现步骤如下：

第一步，将该结点上的所有数据样本按照连续属性的具体数值由小到大进行排序，得到属性值的取值序列为$\{A1c, A2c,\cdots, Atc\}$。

第二步，在取值序列$[A1c, Atc]$中生成$t-1$个分割点。第i（$0<i<t$）个分割点的取值设置为

$$Vi=(Aic+A(i+1)c)/2,$$

其可以将该节点上的数据集划分为2个子集。

第三步，计算$t-1$个分割点所对应分类的信息增益比，从中选择最大信息增益率所对应的分割点作为属性Ac分类的分割点。

1.3 采用悲观错误算法剪枝

悲观错误剪枝（pessimistic error pruning, PEP）算法是一种后剪枝算法。该算法先生成与训练数据完全一致的决策树，然后由底向上搜索，从树最底层的内部结点将符合修剪规则的结点剪掉。悲观错误剪枝在实际应用中表现出了较高的精度，其不需要分离的剪枝数据集，这对于事例较少的问题非常有利。该方法中使用的公式[2]如下，

$$Pr\left[\frac{f-q}{\sqrt{q(1-q)/N}}>z\right]=c, \qquad (3)$$

式中：c表示置信度，默认值为0.25，是C4.5算法的一个输入参数；

q为真实的误差率；

N为实例数量；

z为对应于置信度c的标准差；

f为观察到的误差率，即$f=E/N$。

通过式（3）计算出真实误差率q的一个置信度上限，再根据上限，对该节点误差率e做一个悲观的估计[3]，即

$$e=\frac{f+\frac{Z^2}{2N}+Z\sqrt{\frac{f}{N}-\frac{f^2}{N}+\frac{Z^2}{4N^2}}}{1+\frac{Z^2}{N}}。 \qquad (4)$$

2 数据预处理

本文所有数据来源于株洲某健身俱乐部，而这些原始的数据不能直接用于数据挖掘，因为其存在的一些噪声、缺失数据和不一致性数据会给数据挖掘的结果产生较大影响。因此，对这些数据进行数据挖掘前，必须先进行预处理。数据预处理包含数据集成、数据归约、数据清理和数据变换5种方法。

2.1 数据归约

数据归约技术用于产生数据的归约表示，选择与数据挖掘应用相关的数据，使数据的范围减小，以达到用最小的测量和处理量获得最好的性能。其主要方法包括数据立方体聚集、维归约、数据压缩、数值归约、离散化和概念分层等。本文采用数值规约和离散化处理对2种数据进行归约，即会员健身记录和会员档案信息。所有会员健身记录截止于2012年8月28日，得到共60 000条记录，并将其按会员健身登记时间降序排列。只保留会员卡在有效期内的会

员健身记录，因此，可将 60 000 条记录归约到 30 199 条。根据会员档案表，该俱乐部的会员人数有 2 953 人，由于只针对会员卡在有效期内的会员，因此，可以将其归约到 985 人。在这 985 人中，其中有 13 人并没有参加健身，由于这 13 人信息对本次挖掘没有帮助，应予以清除，所以最后保留了有效会员档案信息 972 条。

2.2　数据变换

数据变换是把原始数据转换成适合数据挖掘的形式，包括对数据的汇总和聚集、概化、规范化，还可能需要进行属性的构造。在会员健身记录中增加了星期和时间段属性，如从会员健身登记时间“2012-04-15 16:45:00”中提取星期属性为周日，时间段属性为下午。部分会员健身记录数据集如表 1 所示。

表 1　会员健身记录数据集

Table 1　The member exercise-record data set

编号	登记时间	星期	时间段
1347	2012-03-13T19:11	星期二	晚上
1348	2012-03-14T15:20	星期三	下午
1349	2012-03-14T16:20	星期三	下午
1350	2012-03-14T20:10	星期三	晚上

在会员档案表中，由于分类计数采用的是离散型数值，而年龄是一个连续性数值，因此，需要把连续属性离散化。采用概念分层方法将年龄分为 3 层：30 岁以下为青年，30 岁到 50 岁为中年，50 岁以上为老年，部分会员档案数据如表 2 所示。

表 2　会员档案数据集

Table 2　The member file data set

编号	会员卡类型	身份证号码	性别	年龄段
1940	次卡	430×××19770807××13	男	中年
2443	半年卡	430×××19880807××12	女	青年
2539	半年卡	430×××19873892××2X	女	青年
3092	月卡	430×××19780891××12	女	中年

将以上 2 个表合并汇总处理，部分预处理数据如表 3 所示。

表 3　合并后的预处理数据集

Table 3　The combined preprocessing data set

编号	性别	年龄段	会员卡类型
2043	女	青年	次卡
2212	男	青年	次卡
2931	男	中年	半年卡
3001	男	青年	半年卡

为了方便对数据进行整理分析，本文选用统计产品与服务解决方案（statistical product and service solutions，SPSS）作为统计分析软件，并对数据进行规范化规定，如表 4 所示。

表 4　规范化约定

Table 4　Standardization agreement

属性字段	规范化约定
性别	0：男，1：女。
年龄段	0：青年，1：中年，2：老年。
卡类型	0：年卡；1：百日卡；2：翔龙卡；3：商务会员卡；4：跨年卡；5：半年卡；6：季卡；7：暑假卡；8：520 卡 9：商务 VIP 会员卡；10：金龙卡；11：次卡；12：月卡；13：学期卡；14：周末卡；15：双月卡。
是否正例	正例：用 1 表示，即在非周末白天的健身记录。反例：用 0 表示，即在周末或非周末的晚上的健身记录。
叶子节点	当正例率超过 35% 为 Y，否则为 N。

预处理后的部分数据集如表 5 所示。

表 5　预处理后的待挖掘的数据集

Talbe 5　The preprocessed data set for mining

编号	性别	年龄段	卡类型	是否正例
1897	1	0	11	0
1876	1	0	11	1
2101	0	1	6	1
2131	1	0	5	1

2.3　数据清理

现实世界中的数据通常是“脏的”，数据清理包括：填充空缺值，识别孤立点，去掉原始数据中的噪声和无关数据。在本数据挖掘中，2 个重要的考虑对象是会员的年龄和性别。由于前台操作员有时忘了输入这些数据，或者一些特殊的情况使其空缺值情况较严重。可以按以下步骤处理这些“脏数据”：

第一步，通过身份证号码识别年龄和性别。身份证号码中第 7 到第 10 位数字为出生年份，只需将现在的年份减去该会员的出生年份，就可以得到其年龄。如果身份证号码为 15 位数，最后一位数是双数表示女，单数表示男；如果身份证号码为 18 位数，倒数第二位（第 17 位）是双数表示女，单数表示男。

第二步，通过会员照片人工对会员的年龄和性别予以识别。这种方法不仅耗时而且有错误率的存在，特别是对于年龄段的识别，带有较大的主观性。

第三步，针对少数既没有录入身份证号码，也没有登记会员照片的情况。如健身俱乐部将会员卡赠送给某个单位（株洲电视台或红网）的员工，而这些会员既没有录入身份证号码和性别，也没有登记会员照片。俱乐部当前有效会员为 972 人，其中有 9 条这种＂脏数据＂，可以请健身俱乐部的工作人员予以配合，因为每个会员都会指定一个固定的会籍顾问，从会籍顾问那里可以确定这 9 张会员卡使用人的年龄和性别。

3 建立决策树

利用C4.5算法对经过预处理后的数据进行数据挖掘，即建立决策树。

3.1 计算类别属性的信息量

决策树中每一个非叶子结点对应着一个非类别属性，树枝代表这个属性的值。一个叶子结点代表从树根到叶子结点之间的路径对应的记录所属的类别属性值。类别属性有是否正例，非类别属性包括性别、年龄段、会员卡类型。

会员健身记录有30 199条，其中对于类别属性（C）是否正例，正例有6 019条，反例有24 456条。属性C分割会员健身记录的训练集对应的信息量为

$$S_{\text{SplitInfo}}(C)=-\frac{5\ 996}{30\ 199}\log_2\frac{5\ 996}{30\ 199}-\frac{24\ 203}{30\ 199}\times\log_2\frac{24\ 203}{30\ 199}=0.719\ 0。$$

3.2 计算非类别属性的信息增益率

对非类别属性性别（S）进行计算，共有会员健身记录30 199条，其中男性会员有14 924人（正例有3 176例，反例有1 1748例），女性会员有1 5275人（正例有2 820例，反例有12 455例）。属性S分割会员健身记录的训练集对应的信息量为

$$S_{\text{SplitInfo}}(S)=-\frac{14\ 924}{30\ 199}\log_2\frac{14\ 924}{30\ 199}-\frac{15\ 275}{30\ 199}\times\log_2\frac{15\ 275}{30\ 199}=0.999\ 9。$$

性别条件信息熵为

$$H(C|S)=\frac{14\ 924}{30\ 199}\left(-\frac{3\ 176}{14\ 924}\log_2\frac{3\ 176}{14\ 924}-\frac{11\ 748}{14\ 924}\log_2\frac{11\ 748}{14\ 924}\right)+\frac{15\ 275}{30\ 199}\left(-\frac{2\ 820}{15\ 275}\log_2\frac{2\ 820}{15\ 275}-\frac{12\ 455}{15\ 275}\log_2\frac{12\ 455}{15\ 275}\right)=0.718\ 1。$$

性别信息增益率为

$$G_{\text{Gain\_Ratio}}(S)=\frac{S_{\text{SplitInfo}}(C)-H(C|S)}{H(S)}=\frac{0.719\ 0-0.718\ 1}{0.999\ 9}=0.000\ 9。$$

同理依次计算出会员卡类型和年龄段的信息增益率，即

$G_{\text{Gain\_Ratio}}$(会员卡类型)=0.000 769，

$G_{\text{Gain\_Ratio}}$(年龄段)=0.007 21。

3.3 生成根节点

通过上面计算得到的性别、会员卡类型和年龄段属性信息增益率可以看出，年龄段属性的信息增益率最大，即年龄段对分类的影响最大，所以选择年龄段属性作为树的根结点属性。生成根结点如图1所示。

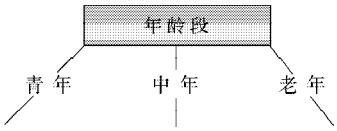

图1 决策树根节点

Fig. 1 The root node of decision tree

3.4 生成决策树

按照C4.5算法递归计算，一层层的找出结点上的属性。树的增长方式采用CHAID算法，CHAID算法提供了一种在多个自变量中自动搜索产生最大差异的变量方案。由C4.5算法生成的一个未剪枝的决策树模型，如表6所示。

表6 未剪枝的决策树模型

Table 6 The model of decision tree without pruning

增长方式	CHAID
因变量	是否正例
自变量	性别，年龄段，会员卡类型
验证最大树深度	3
节点数	29
终端节点数深度	3

按悲观错误剪枝方法对该决策树进行剪枝[4]，最终生成二叉树，如图2所示。

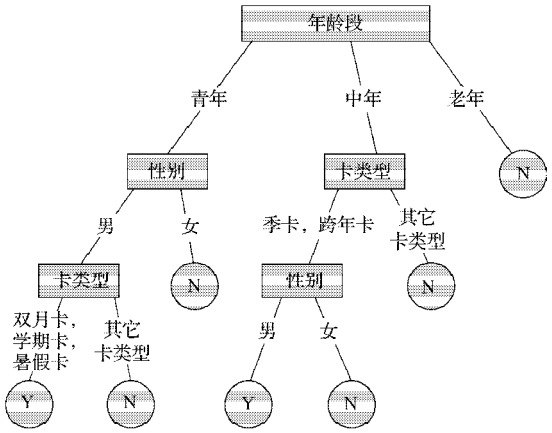

图2 按照C4.5算法生成的二叉树

Fig. 2 The two fork tree based on C4.5 algorithm

4 实验结果分析

从决策树中提取分类规则如下：

if age=0 and sex=0 and (card=15 or card=13 or card=7) then classes=y

if age=1 and (card=6 or card=4) and sex=0 then classes=y。

系统当前有效会员为 972 人，将此规则应用到原会员卡管理系统中，分类出非周末白天健身的会员为 139 人。健身俱乐部管理者可以针对这部分会员制定更适合他们的健身课程。会员信息查询界面见图 3。

图 3　将 C4.5 算法应用到会员管理系统中

Fig. 3　The application of C4.5 algorithm to the member management system

5　结语

将数据挖掘 C4.5 决策树算法应用到健身俱乐部的会员分类中，挖掘出非周末白天参与健身的会员特征，提取出分类规则并应用到当前会员管理系统中，使健身俱乐部管理者可以针对这部分会员制定健身教练课程的营销方案，有针对性地发送短信给特定分类的会员，从而节约运营成本。本课题组下一步工作的重点是全面分析健身俱乐部营业额季节性周期变化规律与会员特征之间的关联规则，以提高健身俱乐部的竞争力和市场占有率。

参考文献：

[1] 云玉屏. 基于 C4.5 算法的数据挖掘应用研究[D]. 哈尔滨：哈尔滨理工大学，2008.
Yun Yuping. Application and Research of Data Mining Based on C4.5 Algorithm[D]. Harbin：Harbin University of Science and Technology，2008.

[2] 李　会，胡笑梅. 决策树中 ID3 算法与 C4.5 算法分析与比较[J]. 水电能源科学，2008，26(2)：129-134.
Li Hui，Hu Xiaomei. Analysis and Comparison between ID3 Algorithm and C4.5 Algorithm in Decision[J]. Water Resources and Power，2008，26(2)：129-134.

[3] 王斌会. 数据挖掘技术及其应用现状[J]. 统计与决策，2006 (10)：122-124.
Wang Binhui. Data Mining Technology and Its Application[J], Statistics and Decision, 2006(10)：122-124.

[4] 吴　陈，林炎钟. C4.5 算法在高校教师评价中的应用研究[J]. 信息技术，2011 (1)：133-136.
Wu Chen，Lin Yanzhong. Application Research on C4.5 Algorithm in High School Teacher Valuation[J]. Information Technology, 2011(1)：133-136.

[5] 胡海斌，邱　明，姜青山，等. 一种基于数据继承关系的C4.5关系分类优化算法[J]. 计算机研究与发展，2009，46（增刊 2）：491-495.
Hu Haibin，Qiu Ming，Jiang Qingshan，et al. An Improved Classification Algorithm of C4.5 Based on Data's Inheritance Relationship[J]. Journal of Computer Research and Development, 2009，46(S2)：491-495.

（责任编辑：邓　彬）