数据挖掘技术决策树分类算法分析、比较与实验

马俊宏
（晋中学院，晋中 030600）

摘 要： 近些年来，互联网迅速发展，数据量每年都以惊人的幅度提升，人们的生活、政府的管理都和电子信息设备息息相关，特别是电子商务和科学实验数据库的迅速壮大，为我们带来了海量的数据。这些海量的数据中，往往蕴藏非常多有价值的记录和信息，等待着人们去挖掘，人们希望将这些信息分离提取出来进行更高程度的分析和统计，以便为我们所取用。而目前大部分数据库系统仅仅可以实现数据的增、删、改、查，很难找到大数据之间所蕴含的规则和关系，比较缺乏挖掘数据内部价值的有效方法，较难通过数据的维度去探索和发现、预测未来的趋势。本文通过对数据挖掘技术中决策树的分类算法做出实验分析，进行比较，给出合理的分析建议。

关键词： 数据挖掘；决策树；ID3算法

中图分类号：TP31 文献标志码：A 文章编号：1004-8626(2017)07-0159-03

Analysis, Comparison and Experiment of Classification Algorithm of Decision Tree in Data Mining Technology

Ma Junhong
(Jinzhong University, Jinzhong 030600, China)

Abstract: In recent years, Internet has developed rapidly. The amount of data increases at an alarming range every year. People's life and management of government are closely related to electronic information equipment. In particular, the rapid growth of e-commerce and scientific experiment database has brought us huge amounts of data. A lot of valuable records and information are stored in vast amounts of data, waiting for mining. People want to separate and extract these information for a higher level of analysis and statistics so that we can use them. However, at present, most database systems can only achieve data increasing, deleting, changing and checking, in which it is hard to find the rules and relationships between big data without effective ways to tap the internal value of data and it is difficult to explore, discover and predict future trend through the dimension of data. This paper makes an experimental analysis of the classification algorithm of decision tree in data mining technology, compares and gives reasonable analysis suggestions.

Keywords: data mining; decision tree; ID3 algorithm

一、绪论

（一）数据挖掘

在海量数据中提取有价值的信息和知识被我们称之为数据挖掘技术。在海量数据库、云端服务器、数据仓储等存储媒介里面都存放着大量的数据信息，我们可以在这些存储媒介当中去探寻有价值的数据，深入地分析和挖掘数据中的内在价值。帮助决策者找寻数据与数据之间可能存在的潜在关联结构，及时有效的发现可能被忽略和遗忘的要点。通常来说，这些数据信息对未来趋势的行为判断有着重要的作用，从而引导决策者做出正确的判断和最优的决策。因此人们发明的决策树分类算法，来帮助人们更好的挖掘数据中有价值的信息。决策树分类算法的挖掘过程可能要多次循环往复螺旋递进，直至达到我们想要的结果。（见图1）

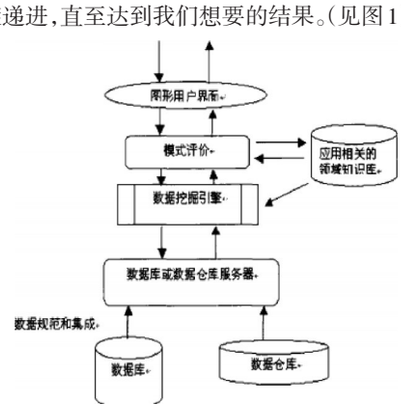

图1 数据挖掘系统架构图

收稿日期：2017-10-22
基金项目：《大数据工作室》，晋中学院“1331工程”重点创新团队建设计划资助科研课题

（二）数据挖掘分类算法的意义

目前来看，数据挖掘在实际应用中有着重要的作用和意义，数据挖掘技术可以运用于很多场合。比如在股票金融市场中，可以对股票的历史交易信息数据进行分析和预测，并对其涨跌走势做出比较准确的判断；再比如在天气预报的过程中，对空气各类成分以及近半个月的数据进行收集、处理和分析，可以对天气预报做出比较准确的合理预测；在产品的销售系统中，已存原始数据库信息，现在假定有新的客户添加进数据库中，我们想讲广告促销信息分发给顾客。如果每一位顾客都通知，这势必成本较大，耗费较多，此时通过数据挖掘技术，找到那些比较有意向购买的顾客，向他们推送广告，可以大大节约了时间和金钱费用，促进的成交量，为商家带来更大的经济效益。数据挖掘技术其实就是一种决策支持的过程，是对数据进行深层次的数据分析方法。在平常生活中，可以将数据挖掘技术应用于方方面面，对促进社会的进步和发展有着很大的帮助。因此对决策树分类算法的相关研究有着较高的实用价值和研究价值。

二、决策树分类算法相关知识

（一）决策树的介绍

决策树（Decision Tree，DT）是一种常用的分类方法，适用于解决各种的分类问题。它通过将数据集进行分类、聚类和预测建模，将一个整体的大问题逐个逐个分解成每个子集小问题，再逐个一已解决子集问题，提高解决问题的效率。通常我们需要构建一个决策树来对分类过程进行建模比较。

（二）决策树基本原理

1948年，美国数学家克劳德-艾而德伍-香农（Claude Elwood Shannon）创建了信息论，用来解决在信息传递过程中的不确定性等问题。在信息论的基础上，决策树运用技术发展壮大。它通过数学的方法度量分析信息数据，通过自定义不同的符号情况，来描绘信息量的大小。其中包括一系列相关概念描述，以下为具体展示：

（1）自信息量。设连续发出的信号为X1、X2.....Xn为发出的信号，直到接收Xi信号，把不确定性的信号标识为I（Xi），即（2.1）

$$I(X_i) = -\log_2 P(X_i) \qquad (2.1)$$

其中P（Xi）表示信源发出Xi的概率。

（2）信息熵。再通过信息熵来度量信号源X的不确定性，即（2.2）

$$H(X) = -\sum_{i=1}^{n} P(X_i)\log_2 P(X_i) \qquad (2.2)$$

其中X为信号源，i为任意可能的符号数。

（3）条件熵。设信号源X和Y不是相互独立的，则用条件熵H（X/Y）来度量整体的不确定性。设X对应的信号源为Xi，Y对应的信号源为Yj，则有：

$$H(X/Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P(X_iY_j)\log_2 P(X_i/Y_j) \qquad (2.3)$$

（4）平均互信息量。信号源X和Y之间的相互关系：

$$I(X,Y)=H(X)-H(X/Y) \qquad (2.4)$$

依据信息论，设S为整个样本数据整体集合，其中包含n类训练数据集，每类有Si个实例，则把它们分类所需要的信息量I用如下公式2.5表示为：

$$I(S_1,S_2,...,S_n) = -\sum_{i=1}^{n} P_i \log_2(P_i) \qquad (2.5)$$

由此，我们可以得到数据样本为S的包含N类的数据集，为了使下一步的工作尽可能尽量的减小，要求每一次都选择信息增益最大的属性作为决策树的节点，并对属性进行划分建立分枝，依据此思想划分数据样本集。

三、决策树ID3算法分析

（一）决策树模型的建立

以下我们通过一个具体示例来演示经典ID3算法的整个构建过程。我们采用来自All Electronics顾客数据库数据元组训练集。

利用ID3算法对数据集合进行决策树模型的建立，对顾客进行分类，整个计算过程如下：

1.计算给定样本集的信息熵，我们使用以下公式进行计算：$I(s_1,s_2,.....,s_m) = -\sum_{i=1}^{m} P_i \log_2(P_i)$

所以

$$I = I(9,5) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940$$

2.计算每个属性的信息增益

（1）需要确定属性age的每个样本值yes和no的分布。

*如果age="<=30"，则p1=2（有2个yes），n1=3（有3个no），

由公式计算可知：

$$I(p_1,n_1) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971;$$

*如果age="31…40"，则p2=4（有4个yes），n2=0（0个no），由公式计算知：I（p2, n2）=0;

*如果age=">40"，则p3=3（有3个yes），n3=2（2个no），由公式计算可知：I（p3, n3）=0.971;

（2）对于属性income，需要知道income的每个样本值yes和no的分布。

*如果income="high"，则p1=2（此时类别为yes的个数），n1=2（此时类别为no的个数），由公式计算可知：

$$I(p_1,n_1) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$$

*如果income="medium"，则p2=4（有4个yes），n2=2（有2个no），由公式计算可知：I（p2, n2）=0.148;

*如果income="low"，则p3=3（有3个yes），n3=1（有1个no），由公式计算可知：I（p3, n3）=0.279;

（3）生成决策树的根和分枝。如下图2所示，我们可以从图中看出当age为31-40时，节点所对应的类别均为yes值，所以此时该节点的I（p2,n2）节点的信息熵为O,而<=30的属性和>40的属性都还有两个类别，所以要对它们进一步划分。

（4）依照上文所述的算法原理过程，对整体训练数据集进行递归分解，按照数据信息不同属性分为不同类别，最终

（下转第186页）

五、结语

综上所述;大学物理实验虚拟考核(无纸化)打破了传统实验考核的局限,促进了学生课堂积极实验和课外模拟仿真,通过对实体实验和仿真实验资源的整合,将学生能力培养延伸到课外,极大地缓解了高等院校人才培养方案中《基础物理实验》课时少与学生能力培养所需学时数多之间的矛盾,同时也规范了实验教学管理,达到了"以考促教,以考促学,以考促管"的目的。学生通过虚拟考试,对实验课程重视并产生兴趣,教师也通过考核模式的改进而不断调整和提升教学效果。同时,我们也发现一些虚拟考核存在的若干问题,包括操作给分过于模式化,系统严格按照内部程序制定的判断原则给分,部分实验存在不同的操作(连线)方法,往往系统不能正确给予判别;硬件问题,系统依托于服务器运行,往往在考试前和考试中,由于学生大量登陆而导致服务器不定时的崩溃,给考核带来不必要的麻烦。

仿真题库也需要及时更新,否则也出现学生背题的情况。故此,我们要继续总结发现实验虚拟考核的经验和问题,不断改进考核手段,完善配套设施,最终实现大学物理实验整体教学改革的实质性提升。

参考文献:

[1]张道清,梁枫,肖世发.普通物理实验考核方式改革的研究[J].大学物理实验,2017,30(2):129-131.

[2]时阳光,张广斌.大学物理实验考核方式的探索与实践[J].物理与工程,2016(1):241-243.

[3]洪炜宁,钱良存,郭守月,刘家菊,黎珉.农业院校大学物理实验教学改革与实践[J].牡丹江师范学院学报(自然科学版),2011,(4):62-63.

[4]张春平,初建崇,胡慧.大学物理实验多元化考核方式的探索[J].实验技术与管理,2015,32(4):223-225.

[5]张博,毛巍威.《大学物理实验》课程考试改革的探索[J].实验科学与技术,2012,10(6):38-39.

[6]余善好,孙权海,吴诗芬.大学物理实验考核模式探索[J].电脑知识与技术:学术交流,2016,12(1Z):147-149.

[7]钱良存,朱德泉,洪炜宁,刘家菊,黎珉.农业院校大学物理实验教学评考结合机制的实践[J].安徽农学通报,2016,22(13):154-156.

[8]范志强,彭金池,唐贵平.基于网络教学平台的大学物理实验教学改革与实践[J].教育教学论坛,2017(31):99-100.

[9]郭晓春.虚拟仿真技术对大学物理实验教学的影响[J].科教导刊,2017,2(1):115-116.

[10]郭文阁,陈海霞,王玮.仿真实验在大学物理实验教学应用探讨[J].大学物理实验,2015,28.112(3):124-127.

[11]岑铬锋,胡君辉,李丹.大学物理实验虚拟系统设计与交互式教学的实现[J].实验科学与技术,2011,09(5):186-189.

[12]张林.虚拟仿真技术对大学物理实验教学影响的探究[J].大学物理实验,2015,28. 110(1):116-118.

[13]张林,濮兴庭,于莉莉,等.大学物理仿真实验教学的探索与实践[J].广西物理,2012(4):50-52.

[14]王艳华.虚拟仿真实验——大学物理实验教学多元化的平台[J].课程教育研究,2014(32):188-189.

[15]张明,李良荣.在大学物理教学中引入仿真测试技术提高教学效果[J].大学物理,2011,30(6):32-34.

[16]宁锌,庞玮,陈峻.大学物理仿真实验在教学实践中的应用[J].大学物理实验,2012(3):110-111.

(责任编辑:李东升)

(上接第160页)

建成决策树的分类模型,得到决策树的理想化模型。

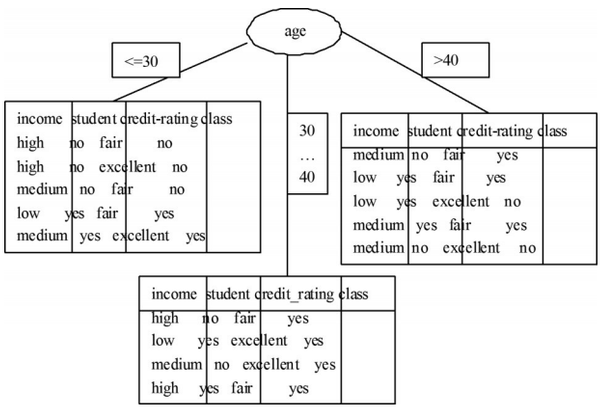

图2 age分枝属性详细图

四、结语

综上所述;在这个信息化的时代,处理大量混乱而又复杂的数据的一个很好的方法是分类,在分类技术的发展过程中,几个流行的技术分别是:神经网络、遗传算法、贝叶斯分类、决策树等。决策树算法理论清晰,效果直观,更易被读者所理解,能够较好的显示出数据之间的关联和内在联系,具有不错的分类预测能力。因此对决策树算法的研究有着重要的研究价值和实际意义。

参考文献

[1] 毛国君,段立娟,王实,等.数据挖掘原理与算法[M].北京:清华大学出版社,2005

[2] Jiawei Han,Micheline Kamber著.范明,孟小峰译.数据挖掘概念与技术[M].北京:机械工业出版社,2001

[3] 美Mehmed Kantardzic著.数据挖掘—— 概念、模型、方法和算法[M].闪四清,陈茵,程雁,等译.北京:清华大学出版社,2003

[4] 张维东.利用决策树进行数据挖掘中的信息熵计算[J].计算机工程,2001(3):66-68.

[5] 王大玲.基于概念层次树的数据挖掘算法的研究与实现[J].计算机科学,2001.2(2):63-66.

[6] 唐华松.数据挖掘中决策树算法的探讨[J].计算机应用研究,2001(8):36-40.

[7] 许兆新.决策支持系统相关技术综述[J].计算机应用研究,2001(2):22-26.

[8] 王熙照.决策树简化(剪切)方法综述[J].计算机工程与应用,2004(40):32-35.

[9] 胡江洪.基于决策树的分类算法研究[J].计算机工程与应用,2005(27):66-69.

(责任编辑:李东升)